

ALAGAPPA UNIVERSITY



[Accredited with 'A+' Grade by NAAC (CGPA:3.64) in the Third Cycle
and Graded as Category–I University by MHRD-UGC]

(A State University Established by the Government of Tamil Nadu)

KARAIKUDI – 630 003

Directorate of Distance Education

M.Sc. [Zoology]

IV - Semester

350 43

BIOPHYSICS, BIOSTATISTICS AND BIOINFORMATICS

Author:
Late K K Sharma, Department of Chemistry, Zakir Husain College, University of Delhi
L K Sharma, Former Associate Professor, Department of Chemistry, ARSD College, University of Delhi
Units: (1.6, 2.8)
Dr. J S Chandan, Prof. Medgar Evers College, City University of New York, New York.
Units: (5, 6.5, 7)
Dr. Pradeep Kumar, Research Associate-III, Department of Pediatrics, Army Hospital Research & Referral, New Delhi
Unit: (13.0-13.3)
Vikas Publishing House, Units: (1.0-1.5, 1.7-1.11, 2.0-2.2, 2.3-2.7, 2.9-2.13, 3, 4, 6.0-6.4, 6.6-6.10, 8, 9, 10, 11, 12, 13.4-13.9, 14)



Vikas® is the registered trademark of Vikas® Publishing House Pvt. Ltd.

VIKAS® PUBLISHING HOUSE PVT. LTD.
E-28, Sector-8, Noida - 201301 (UP)
Phone: 0120-4078900 • Fax: 0120-4078999
Regd. Office: A-27, 2nd Floor, Mohan Co-operative Industrial Estate, New Delhi 1100 44
• Website: www.vikaspublishing.com • Email: helpline@vikaspublishing.com

Work Order No.AU/DDE/DE12-16/Printing of Course Material/2020 Dated 28.02.2020, Copies - 500

SYLLABI-BOOK MAPPING TABLE

Biophysics, Biostatistics and Bioinformatics

Syllabi	Mapping in Book
BLOCK - I: BIOPHYSICS	
Unit I Introduction to Biophysics: Structure and Properties of Atoms and Molecules-Chemical Bonds - Types and Properties; Polymerization of Organic Molecules.	**Unit 1:** Introduction to Biophysics **(Pages 1-37)**
Unit II Laws of Thermodynamics - Principle and Application; Bio-Energetics - Coupling of Chemical Reactions - Redox Potential - NADP/NADPH and Free Energy.	**Unit 2:** Thermodynamics **(Pages 38-75)**
Unit III Natural Radiations - Properties of Light, Absorption of Light, Energy States of Atoms, Spin Property of Electrons- Ground State and Excited State of Atoms and Bio-Molecules - Their Effects.	**Unit 3:** Natural Radiations **(Pages 76-98)**
Unit IV Spectroscopy- Principle and Applications; Delayed Effects of Radiation and Measurement of Radio Activity - Geiger Muller Counter - Isotopes as Tracers - Autoradiography.	**Unit 4:** Spectroscopy **(Pages 99-119)**
BLOCK - II: BIOSTATISTICS	
Unit V Definition and Scope of Biostatistics - Collection of Data - Primary and Secondary Data.	**Unit 5:** Definition and Scope of Biostatistics **(Pages 120-143)**
Unit VI Types of Sampling: Random and Stratified Random Sampling, Types of Variables: Continuous and Discontinuous Variables, Qualitative and Quantitative Variables.	**Unit 6:** Types of Sampling **(Pages 144-152)**
Unit VII Presentation of Data: Line and Bar Diagram, Histogram, Polygon and Pie Diagram.	**Unit 7:** Presentation of Data **(Pages 153-166)**
BLOCK - III: MEASURES OF CENTRAL TENDENCY AND MEASURE OF DISPERSION	
Unit VIII Mean, Median and Mode - Dispersion: Range, Variance, SD, SE and CV.	**Unit 8:** Mean, Median and Mode **(Pages 167-184)**
Unit IX Probability and Hypothesis Testing- Normal Distribution, Confidence Interval and P Value.	**Unit 9:** Probability and Hypothesis Testing **(Pages 185-215)**
Unit X Common Statistical Tools: Chi-Square, 't' Test, ANOVA, Correlation and Regression Analysis; Statistical Packages.	**Unit 10:** Common Statistical Tools **(Pages 216-246)**

BLOCK - IV: BIOINFORMATICS

Unit XI

Introduction to Bioinformatics, Medical - Informatics, Cheminformatics and Pharmacoinformatics.

Unit 11: Introduction to Bioinformatics
(Pages 247-266)

Unit XII

Current Researches in Bioinformatics. Applications of Bioinformatics in Cancer Detection, Drug Targets.

Unit 12: Current Researches in Bioinformatics
(Pages 267-278)

Unit - XIII

Animal Genome Diversity - Introduction to DNA and Protein Sequence Analysis - Introduction and Concepts to Biological Databases.

Unit 13: Animal Genome Diversity
(Pages 279-307)

Unit - XIV

Phylogenetic Analysis using PHYLIP, ClustalW.

Unit 14: Phylogenetic Analysis
(Pages 308-336)

CONTENTS

BLOCK IV: BIOINFORMATICS

INTRODUCTION

NOTES

Biophysics is an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena. Biophysics covers all scales of biological organization, from molecular to organismic and populations. Biophysical research shares significant overlap with biochemistry, molecular biology, physical chemistry, physiology, nanotechnology, bioengineering, computational biology, biomechanics, developmental biology and systems biology. The term biophysics was originally introduced by Karl Pearson in 1892. The term biophysics is also regularly used in academia to indicate the study of the physical quantities (e.g., electric current, temperature, stress, entropy) in biological systems. Other biological sciences also perform research on the biophysical properties of living organisms including molecular biology, cell biology, chemical biology, and biochemistry.

Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results. Biostatistical modeling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools. Gregor Mendel started the genetics studies investigating genetics segregation patterns in families of peas and used statistics to explain the collected data.

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data. Bioinformatics includes biological studies that use computer programming as part of their methodology, as well as a specific analysis 'Pipelines' that are repeatedly used, particularly in the field of genomics. Common uses of bioinformatics include the identification of candidates genes and Single Nucleotide Polymorphisms (SNPs). Often, such identification is made with the aim of better understanding the genetic basis of disease, unique adaptations, desirable properties (esp. in agricultural species), or differences between populations. In a less formal way, bioinformatics also tries to understand the organizational principles within nucleic acid and protein sequences, called proteomics.

This book, *Biophysics, Biostatistics and Bioinformatics*, is divided into four blocks, which are further subdivided into fourteen units. This book provides a basic understanding of the subject and helps to grasp its fundamentals. In a

NOTES

nutshell, it explains various aspects, such as biophysics, structure and properties of atoms and molecules, chemical bonds, polymerization of organic molecules, laws of thermodynamics, bio-energetics, coupling of chemical reactions, NADP/ NADPH, natural radiations, properties of light, energy states of atoms, spin property of electrons, spectroscopy, delayed effects of radiation and measurement of radio activity, Geiger Muller counter, isotopes as tracers, autoradiography, scope of biostatistics, collection of data - primary and secondary data, types of sampling, types of variables, presentation of data, line and bar diagram, histogram, polygon and pie diagram, mean, median and mode; dispersion: range, variance, SD, SE and CV; probability and hypothesis testing, normal distribution, confidence interval and P value, Chi-square, '*t*' test, ANOVA, correlation and regression analysis, statistical packages, bioinformatics, medical - informatics, cheminformatics and pharmacoinformatics, current researches in bioinformatics, applications of bioinformatics in cancer detection, drug targets, animal genome, DNA and protein sequence analysis, biological databases, phylogenetic analysis using PHYLIP, ClustalW.

The book follows the Self-Instructional Mode (SIM) wherein each unit begins with an 'Introduction' to the topic. The 'Objectives' are then outlined before going on to the presentation of the detailed content in a simple and structured format. 'Check Your Progress' questions are provided at regular intervals to test the student's understanding of the subject. 'Answers to Check Your Progress Questions', a 'Summary', a list of 'Key Words', and a set of 'Self-Assessment Questions and Exercises' are provided at the end of each unit for effective recapitulation.

BLOCK - I
BIOPHYSICS

NOTES

UNIT 1 INTRODUCTION TO BIOPHYSICS

Structure

1.0 INTRODUCTION

Biophysics is an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena. Biophysics covers all scales of biological organisation, from molecular to organismic and populations. Biophysical examination parts of the significant overlap with biochemistry, molecular biology, physical chemistry, physiology, nanotechnology, bioengineering, computational biology, biomechanics, developmental biology and systems biology. The term biophysics was originally introduced by Karl Pearson in 1892.

Atoms consist of three basic particles: protons, electrons, and neutrons. The nucleus (centre) of the atom contains the protons (positively charged) and the neutrons (no charge). The outermost regions of the atom are called electron shells and contain the electrons (negatively charged). Atoms have different properties based on the arrangement and number of their basic particles. A molecule is an electrically neutral group of two or more atoms held together by chemical bonds. Molecules are distinguished from ions by their lack of electrical charge. Molecular properties include the chemical properties, physical properties, and structural properties of molecules, including drugs. Molecular properties typically do not include pharmacological or biological properties of a chemical compound.

NOTES

A chemical bond is a lasting attraction between atoms, ions or molecules that enables the formation of chemical compounds. The bond may result from the electrostatic force of attraction between oppositely charged ions as in ionic bonds or through the sharing of electrons as in covalent bonds. The strength of chemical bonds varies considerably; there are 'Strong Bonds' or 'Primary Bonds', such as covalent, ionic and metallic bonds, and 'Weak Bonds' or 'Secondary Bonds', such as dipole–dipole interactions, the London dispersion force and hydrogen bonding.

In polymer chemistry, polymerisation, is a process of reacting monomer molecules together in a chemical reaction to form polymer chains or three-dimensional networks. In chemical compounds, polymerisation can occur via a variety of reaction mechanisms that vary in complexity due to the functional groups present in the reactants and their inherent steric effects. In more straightforward polymerisations, alkenes form polymers through relatively simple radical reactions; in contrast, reactions involving substitution at a carbonyl group require more complex synthesis due to the way in which reactants polymerise. Alkanes can also be polymerised, but only with the help of strong acids.

In this unit, you will study about the introduction to biophysics, structure and properties of atom and molecules, chemical bonds, types and properties, polymerisation of organic molecules.

1.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the basics of biophysics
- Analyse the structure and properties of atom and molecules
- Understand the chemical bond
- Discuss the polymerisation of organic molecules

1.2 BASICS CONCEPTS OF BIOPHYSICS

Biophysics is the field that applies the theories and methods of physics to understand how biological systems work. Biophysics has been critical to understanding the mechanics of how the molecules of life are made, how different parts of a cell move and function, and how complex systems in our bodies—the brain, circulation, immune system, and others— work. Biophysics is, therefore, an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena. Biophysics covers all scales of biological organisation, from molecular to organismic and populations. Biophysical research shares significant overlap with biochemistry, molecular biology, physical chemistry, physiology,

nanotechnology, bioengineering, computational biology, biomechanics, developmental biology and systems biology.

NOTES

The term biophysics was originally introduced by Karl Pearson in 1892. Ambiguously, the term biophysics is also regularly used in academia to indicate the study of the physical quantities (e.g., Electric Current, Temperature, Stress, Entropy) in biological systems, which is, by definition, performed by physiology. However, other biological sciences also perform research on the biophysical properties of living organisms including molecular biology, cell biology, chemical biology, and biochemistry. Biophysics, also known as biological physics, is an interdisciplinary science that applies the principles of physics and chemistry and the methods of mathematical analysis and computer modeling to understand how the mechanisms of biological systems work.

Molecular biophysics typically statements biological questions similar to those in biochemistry and molecular biology, seeking to find the physical underpinnings of biomolecular phenomena. Scientists in this field conduct research concerned with understanding the interactions between the various systems of a cell, including the interactions between DeoxyriboNucleic Acid (DNA), RiboNucleic Acid (RNA) and protein biosynthesis, as well as how these interactions are regulated.

Fluorescent imaging techniques, as well as Electron Microscopy, X-ray Crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, Atomic Force Microscopy (AFM) and Small-Angle Scattering (SAS) both with X-rays and Neutrons [Small-Angle Neutron Scattering (SANS)/ Small-Angle X-ray Scattering (SAXS)] are often used to visualise structures of biological significance. Protein dynamics can be observed by neutron spin echo spectroscopy. Conformational change in structure can be measured using techniques, such as dual polarisation interferometry, circular dichroism, SAXS and SANS. Direct manipulation of molecules using optical tweezers or AFM, can also be used to monitor biological events where forces and distances are at the nanoscale. Molecular biophysicists often consider complex biological events as systems of interacting entities which can be understood, for example through statistical mechanics, thermodynamics and chemical kinetics. By drawing knowledge and experimental techniques from a wide variety of disciplines, biophysicists are often able to directly observe, model or even manipulate the structures and interactions of individual molecules or complexes of molecules.

In addition to traditional (i.e., molecular and cellular) biophysical topics like structural biology or enzyme kinetics, modern biophysics encompasses an extraordinarily broad range of research, from bioelectronics to quantum biology involving both experimental and theoretical tools. It is becoming increasingly common for biophysicists to apply the models and experimental techniques derived from physics, as well as mathematics and statistics, to larger systems such as tissues, organs, populations and ecosystems. Biophysical models are used extensively in

the study of electrical conduction in single neurons, as well as neural circuit analysis in both tissue and whole brain.

Medical physics, a branch of biophysics, is any application of physics to medicine or healthcare, ranging from radiology to microscopy and nanomedicine. For example, physicist Richard Feynman theorised about the future of nanomedicine. He wrote about the idea of a medical use for biological machines. Feynman and Albert Hibbs suggested that certain repair machines might one day be reduced in size to the point that it would be possible to (as Feynman put it) "*Swallow the Doctor*". The idea was discussed in Feynman's 1959 essay, "*There's Plenty of Room at the Bottom*".

Focus as a Subfield

While some colleges and universities have dedicated departments of biophysics, usually at the graduate level, many do not have university-level biophysics departments, instead having groups in related departments, such as biochemistry, cell biology, chemistry, computer science, engineering, mathematics, medicine, molecular biology, neuroscience, pharmacology, physics, and physiology. Depending on the strengths of a department at a university differing emphasis will be given to fields of biophysics.

- **Biology and Molecular Biology** – Gene regulation, single protein dynamics, bioenergetics, patch clamping, biomechanics, virophysics.
- **Structural Biology** – Ångstrom-resolution structures of proteins, nucleic acids, lipids, carbohydrates, and complexes.
- **Biochemistry and Chemistry** – Biomolecular structure, RNA, nucleic acid structure, structure-activity relationships.
- **Computer Science** – Neural networks, biomolecular and drug databases.
- **Computational Chemistry** – Molecular dynamics simulation, molecular docking, quantum chemistry.
- **Bioinformatics** – Sequence alignment, structural alignment, protein structure prediction.
- **Mathematics** – Graph/network theory, population modelling, dynamical systems, phylogenetic.
- **Medicine** – Biophysical research that emphasizes medicine. Medical biophysics is a field closely related to physiology. It explains various aspects and systems of the body from a physical and mathematical perspective. Examples are fluid dynamics of blood flow, gas physics of respiration, radiation in diagnostics/treatment and much more. Biophysics is taught as a preclinical subject in many medical schools, mainly in Europe.
- **Neuroscience** – Studying neural networks experimentally (brain slicing) as well as theoretically (computer models), membrane permittivity, gene therapy, understanding tumours.

- **Pharmacology and Physiology** – Channelomics, biomolecular interactions, cellular membranes, polyketides.
- **Physics** – Negentropy, stochastic processes, and the development of new physical techniques and instrumentation as well as their application.
- **Quantum Biology** – The field of quantum biology applies quantum mechanics to biological objects and problems. Decohered isomers to yield time-dependent base substitutions. These studies imply applications in quantum computing.

NOTES

1.3 STRUCTURE AND PROPERTIES OF ATOM AND MOLECULES

Most of the Universe consists of matter and energy. Matter has mass and occupies space while energy is the capacity to do work. All matter is composed of basic elements that cannot be broken down to substances with different chemical or physical properties. Fundamentally, the elements are substances consisting of one type of atom, for example Carbon atoms make up diamond, and also graphite. Pure (24K) gold is composed of only one type of atom, gold atoms.

Principally, the 'Atoms' are the smallest particle into which an element can be divided. As per the work of Enrico Fermi and his colleagues, it is proved that the atom is divisible, often releasing tremendous energies as in nuclear explosions or (in a controlled fashion in) thermonuclear power plants.

1.3.1 Structure of the Atoms

Typically, an atom is the smallest unit of matter that retains all of the chemical properties of an element. Atoms combine to form molecules, which then interact to form solids, gases, or liquids. For example, water is composed of hydrogen and oxygen atoms that have combined to form water molecules. Several biological processes are dedicated to breaking down molecules into their component atoms so that they can be reconstructed into a more useful molecule.

Atomic Particles

Atoms consist of three basic particles: protons, electrons, and neutrons. The nucleus found at the centre in the place an atom whereas the protons contain positively charged and the neutrons have a no charge like a neutral. The outermost regions of the atom are called electron shells and contain the electrons negatively charged. Atoms have different properties based on the arrangement and number of their basic particles.

The Hydrogen Atom (H) contains only one proton, one electron, and no neutrons. This can be determined using the atomic number and the mass number of the element.

NOTES

Atomic Mass

Protons and neutrons have approximately the same mass, about 1.67×10^{-24} grams. Scientists define this amount of mass as one atomic mass unit (amu) or one Dalton. Although similar in mass, protons are positively charged, while neutrons have no charge. Therefore, the number of neutrons in an atom contributes significantly to its mass, but not to its charge.

Electrons are much smaller in mass compare the protons, weighing only 9.11×10^{-28} grams, or 1/1800 is an atomic mass unit. Therefore, they do not contribute much to an element's overall atomic mass. When considering atomic mass, it is customary to ignore the mass of any electrons and calculate the atom's mass based on the number of protons and neutrons alone.

Electrons contribute greatly to the atom's charge, as each electron has a negative charge equal to the positive charge of a proton. Scientists define these charges as '+1' and '–1'. In an uncharged, neutral atom, the number of electrons orbiting the nucleus is equal to the number of protons inside the nucleus. In these atoms, the positive and negative charges cancel each other out, leading to an atom with no net charge.

Volume of Atoms

Accounting for the sizes of protons, neutrons, and electrons, most of the volume of an atom greater than 99 percent, in fact related to the, empty space. Despite all this empty space, solid objects do not just pass through one another. The electrons that surround all atoms are negatively charged and cause atoms to repel one another, preventing atoms from occupying the same space. These intermolecular forces prevent you from falling through an object like your chair.

Atomic Number and Mass Number

The atomic number is the number of protons in an element, while the mass number is the number of protons plus the number of neutrons.

Atomic Number

Neutral atoms of an element contain an equal number of protons and electrons. The number of protons determines an element's atomic number (Z) and distinguishes one element from another. For example, carbon's atomic number (Z) is 6 because it has 6 protons. The number of neutrons can vary to produce isotopes, which are atoms of the same element that have different numbers of neutrons. The number of electrons can also be different in atoms of the same element, thus producing ions (charged atoms). For instance, iron, Fe, can exist in its neutral state, or in the +2 and +3 ionic states.

Mass Number

An element's mass number (A) is the sum of the number of protons and the number of neutrons. The small contribution of mass from electrons is disregarded in calculating the mass number. This approximation of mass can be used to easily

NOTES

calculate how many neutrons an element has by simply subtracting the number of protons from the mass number. Protons and neutrons both weigh about one atomic mass unit or amu. Isotopes of the same element will have the same atomic number but different mass numbers.

Scientists determine the atomic mass by calculating the mean of the mass numbers for its naturally occurring isotopes. Often, the resulting number contains a decimal. For example, the atomic mass of Chlorine (Cl) is 35.45 amu because Chlorine is composed of several isotopes, some (the majority) with an atomic mass of 35 amu (17 protons and 18 neutrons) and some with an atomic mass of 37 amu (17 protons and 20 neutrons).

Given an atomic number (Z) and mass number (A), we can find the number of protons, neutrons, and electrons in a neutral atom. For example, a Lithium tom (Z=3, A=7 amu) contains three protons (found from Z), three electrons (as the number of protons is equal to the number of electrons in an atom), and four neutrons (7 – 3 = 4).

Properties of the Atoms

- The physical world is composed of combinations of various subatomic or fundamental particles. These are the smallest building blocks of matter.
- The atoms consist of two parts. An atomic nucleus and an electron cloud.
- The number of electrons and their arrangement in the electron cloud is responsible for the chemical behaviour of atoms.
- The nuclear properties (atomic mass, nuclear cross-sections) of the element are determined by the number of protons (atomic number) and number of neutrons (neutron number).
- Nuclear stability is a concept that helps to identify the stability of an isotope. To identify the stability of an isotope it is needed to find the ratio of neutrons to protons. To determine the stability of an isotope we can use the ratio neutron/proton (N/Z).
- There are only certain combinations of neutrons and protons, which forms stable nuclei.
- Unstable nuclei must undergo nuclear decay (radioactive decay) to stabilise itself, it is a random and natural process.
- The number of atoms in 1 mole (e.g., 12 grams of carbon) of a substance is equal to the Avogadro's constant, which is equal to 6.022×10^{23} mol^{-1}.

1.3.2 Structure of the Molecules

Molecular biophysics is the study of the physical principles governing biomolecular systems. It seeks to explain biological function in terms of molecular structure, dynamics and organisation, from single molecules to supramolecular structures. Molecular biophysics is a rapidly evolving interdisciplinary form of the area of

NOTES

research that combination concepts in the fields of physics, chemistry, engineering, mathematics and biology. It seeks to understand biomolecular systems and explain biological function in terms of molecular structure, structural organisation, and dynamic behaviour at various levels of complexity (from single molecules to supramolecular structures, viruses and small living systems). This discipline covers topics, such as the measurement of molecular forces, molecular associations, allosteric interactions, Brownian motion, and Cable theory. Additional areas of study can be found on Outline of Biophysics. The discipline has required development of specialised equipment and procedures capable of imaging and manipulating minute living structures, as well as novel experimental approaches. Scientists in this field conduct research concerned with understanding the interactions between the various systems of a cell, including the interactions between DNA, RNA and protein biosynthesis, as well as how these interactions are regulated.

A molecule is an electrically neutral group of two or more atoms held together by chemical bonds. Molecules are distinguished from ions by their lack of electrical charge. In quantum physics, organic chemistry, and biochemistry, the distinction from ions is dropped and molecule is often used when referring to polyatomic ions. According to the kinetic theory of gases, the term molecule is often used for any gaseous particle regardless of its composition. This violates the definition that a molecule contain two or more atoms, since the noble gases are individual atoms. A molecule may be *homonuclear*, that is, it consists of atoms of one chemical element, as with two atoms in the oxygen molecule (O_2) or it may be *heteronuclear*, a chemical compound composed of more than one element, as with water (two hydrogen atoms and one oxygen atom, which made up the Dihydrogen Monoxide (H_2O) molecule).

Atoms and complexes connected by non-covalent interactions, such as hydrogen bonds or ionic bonds, are typically not considered single molecules. Molecules as components of matter are common. They also make up most of the oceans and atmosphere. Most organic substances are molecules. The substances of life are molecules, e.g., proteins. The amino acids are made of the 'Nucleic Acids (DNA and RNA)', sugars, carbohydrates, fats, and vitamins. The nutrient minerals ordinarily are not molecules, e.g., Iron Sulphate.

Bonding

Molecules are held together by either *covalent bonding* or *ionic bonding*. Several types of non-metal elements exist only as molecules in the environment. For example, hydrogen only exists as hydrogen molecule. A molecule of a compound is made out of two or more elements. A *homonuclear* molecule is made out of two or more atoms of a single element.

While some fact state that metallic crystal can be considered a single giant molecule held together by metallic bonding, others point out that metals act very differently than molecules.

NOTES

Covalent Bond: A covalent bond is a chemical bond that involves the sharing of electron pairs between atoms. These electron pairs are termed shared pairs or bonding pairs, and the stable balance of attractive and repulsive forces between atoms, when they share electrons, is termed covalent bonding.

Ionic Bond: Ionic bonding is a type of chemical bond that involves the electrostatic attraction between oppositely charged ions, and is the primary interaction occurring in ionic compounds. The ions are atoms that have lost one or more electrons (cations) and atoms that have gained one or more electrons (anions). The transfer of an electrons is termed electrovalence in contrast to covalence. In the simplest case, the cation is a metal atom and the anion is a non-metal atom, but these ions can be of a more complicated nature, e.g., molecular ions like NH_4^+ or SO_4^{2-}. At normal temperature and pressure, ionic bonding mostly creates solids (or infrequently liquids) without separate identifiable molecules, but the vaporisation/ sublimation such materials does produce small separate molecules where electrons are still transferred fully enough for the bonds to be considered ionic rather than covalent.

Molecular Size

Most molecules are far too small to be seen with the naked eye, although molecules of many polymers can reach macroscopic sizes, including biopolymers, such as DNA. Molecules commonly used as building blocks for organic synthesis have a dimension of a few Angstroms (Å) to several dozen Å, or around one billionth of a metre. Single molecules cannot usually be observed by light, but small molecules and even the outlines of individual atoms may be traced in some circumstances by use of an atomic force microscope. Some of the largest molecules are macromolecules or super molecules. The smallest molecule is the diatomic hydrogen (H_2), with a bond length of 0.74 Å.

Molecular Formulas

A molecular formula consists of the chemical symbols for the constituent elements followed by numeric subscripts describing the number of atoms of each element present in the molecule. Molecular formulas, typically, describe the exact number and type of atoms in a single molecule of a compound. The constituent elements are represented by their chemical symbols, and the number of atoms of each element present in each molecule is shown as a subscript following that element's symbol, for example molecular formula for glucose is $C_6H_{12}O_6$.

Chemical Formula: The chemical formula for a molecule uses one line of chemical element symbols, numbers, and sometimes also other symbols, such as parentheses, dashes, brackets, and plus (+) and minus (“) signs. These are limited to one typographic line of symbols, which may include subscripts and superscripts.

A compound's empirical formula is a very simple type of chemical formula. It is the simplest integer ratio of the chemical elements that constitute it. For example, water is always composed of a 2:1 ratio of hydrogen to oxygen atoms, and Ethanol

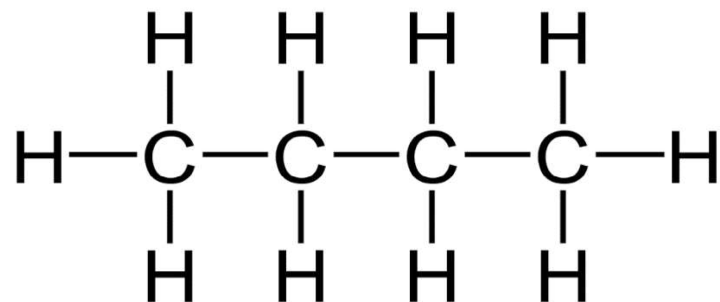
NOTES

(Ethyl Alcohol) is always composed of carbon, hydrogen, and oxygen in a 2:6:1 ratio. However, this does not determine the kind of molecule uniquely dimethyl ether has, for example dimethyl ether has the same ratios as ethanol. Molecules with the same atoms in different arrangements are called isomers. Also carbohydrates, for example, have the same ratio (Carbon: Hydrogen: Oxygen = 1:2:1) but different total numbers of atoms in the molecule.

The molecular formula reflects the exact number of atoms that compose the molecule and so characterises different molecules. However different isomers can have the same atomic composition while being different molecules.

Structural Formula: Molecular formulas contain no information about the arrangement of atoms. Because of this, one molecular formula can describe a number of different chemical structures. A structural formula is used to indicate not only the number of atoms, but also their arrangement in space. A structural formula is not as compact and easy to communicate, but it provides information that the molecular formula does not about the relative positioning of atoms and the bonding between atoms. Compounds that share a chemical formula but have different chemical structures are known as isomers, and they can have quite different physical properties. For molecules with a complicated 3-dimensional structure, especially involving atoms bonded to four different substituents, a simple molecular formula or even semi-structural chemical formula may not be enough to completely specify the molecule. In this case, a graphical type of formula called a structural formula may be required. Structural formulas may in turn be represented with a one-dimensional chemical name, but such chemical nomenclature requires many words and terms which are not part of chemical formulas.

Following is the structural formula of Butane (C_4H_{10}), the chemical structure of butane indicates not only the number of atoms, but also their arrangement in space.



Molecular properties include the chemical properties, physical properties, and structural properties of molecules. Remember that the molecular properties typically do not include pharmacological or biological properties of a chemical compound.

1.4 CHEMICAL BONDS

NOTES

Chemical bonds are forces that hold the atoms together in a molecule. They are a result of strong intramolecular interactions among the atoms of a molecule. The valance (outermost) electrons of the atoms participate in chemical bonds. When two atoms approach each other, these outer electrons start to interact. Although electrons repel each other, they are attracted to the protons within atoms. The interplay of forces results in the formation of bonds between the atoms. The main types of chemical bonds are ionic bond, covalent bond, hydrogen bond, and metallic bond.

A chemical bond is a lasting attraction between atoms, ions or molecules that enables the formation of chemical compounds. The bond may result from the electrostatic force of attraction between oppositely charged ions as in ionic bonds or through the sharing of electrons as in covalent bonds. The strength of chemical bonds varies considerably, there are 'Strong Bonds' or 'Primary Bonds', such as covalent, ionic and metallic bonds, and 'Weak Bonds' or 'Secondary Bonds', such as dipole–dipole interactions, the London dispersion force and hydrogen bonding.

Since opposite charges attract via a simple electromagnetic force, the negatively charged electrons that are orbiting the nucleus and the positively charged protons in the nucleus attract each other. An electron positioned between two nuclei will be attracted to both of them, and the nuclei will be attracted toward electrons in this position. This attraction constitutes the chemical bond. Due to the matter wave nature of electrons and their smaller mass, they must occupied a much larger amount of volume compared with the nuclei, and this volume occupied by the electrons keeps the atomic nuclei in a bond relatively far apart, as compared with the size of the nuclei themselves.

In general, strong chemical bonding is associated with the sharing or transfer of electrons between the participating atoms. The atoms in molecules, crystals, metals and diatomic gases indeed most of the physical environment around us are held together by chemical bonds, which directive the structure and the bulk properties of matter.

All bonds can be explained by quantum theory, but, in practice, simplification rules allow chemists to predict the strength, directionality, and polarity of bonds. The octet rule and Valence Shell Electron Pair Repulsion (VSEPR) theory are two examples. More sophisticated theories are Valence Bond Theory (VBT), which includes orbital hybridisation and resonance, and molecular orbital theory which includes linear combination of atomic orbitals and ligand field theory. Electrostatics are used to describe bond polarities and the effects they have on chemical substances.

NOTES

Because atoms and molecules are three-dimensional, it is difficult to use a single method to indicate orbitals and bonds. In molecular formulas the chemical bonds (binding orbitals) between atoms are indicated in different ways depending on the type of discussion. Sometimes, some details are neglected. For example, in organic chemistry one is sometimes concerned only with the functional group of the molecule. Thus, the molecular formula of Ethanol may be written in conformational form, three-dimensional form, full two-dimensional form (indicating every bond with no three-dimensional directions), compressed two-dimensional form (CH_3–CH_2–OH), by separating the functional group from another part of the molecule (C_2H_5OH), or by its atomic constituents (C_2H_6O), according to what is discussed. Sometimes, even the non-bonding valence shell electrons (with the two-dimensional approximate directions) are marked, e.g., for elemental Carbon (C).

Figure 1.1 illustrates the Lewis dot-style representations of chemical bonds between the Carbon (C), Hydrogen (H), and Oxygen (O).

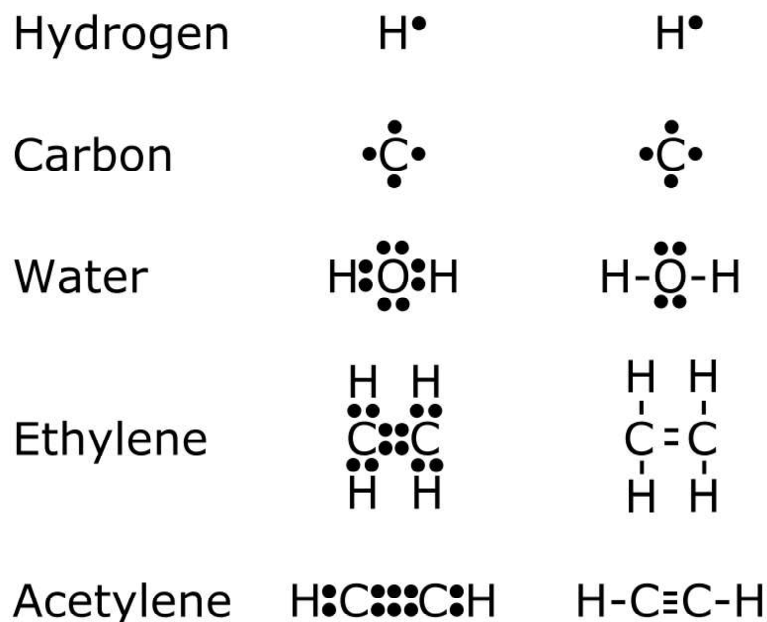


Fig. 1.1 *Lewis Dot-Style Representations of Chemical Bonds between Carbon (C), Hydrogen (H), and Oxygen (O)*

Theories of Chemical Bonding

In the (unrealistic) limit of pure ionic bonding, electrons are perfectly localised on one of the two atoms in the bond. Such bonds can be understood by classical physics. The forces between the atoms are characterised by isotropic continuum electrostatic potentials. Their magnitude is in simple proportion to the charge difference.

Covalent bonds are better understood by Valence Bond (VB) theory or Molecular Orbital (MO) theory. The properties of the atoms involved can be understood using concepts such as oxidation number, formal charge, and electronegativity. The electron density within a bond is not assigned to individual atoms, but is instead delocalised between atoms. In valence bond theory, bonding is conceptualised as being built up from electron pairs that are localised and shared by two atoms via the overlap of atomic orbitals. The concepts of orbital hybridisation and resonance augment this basic notion of the electron pair bond. In molecular orbital theory, bonding is viewed as being delocalised and apportioned in orbitals that extend throughout the molecule and are adapted to its symmetry properties, typically by considering Linear Combinations of Atomic Orbitals (LCAO). Valence bond theory is more chemically intuitive by being spatially localised, allowing attention to be focused on the parts of the molecule undergoing chemical change. In contrast, molecular orbitals are more natural from a quantum mechanical point of view, with orbital energies being physically significant and directly linked to experimental ionisation energies from photoelectron spectroscopy.

Consequently, valence bond theory and molecular orbital theory are often viewed as competing but complementary frameworks that offer different insights into chemical systems. As approaches for electronic structure theory, both MO and VB methods can give approximations to any desired level of accuracy, at least in principle. However, at lower levels, the approximations differ, and one approach may be better suited for computations involving a particular system or property than the other.

Unlike the spherically symmetrical Coulombic forces in pure ionic bonds, covalent bonds are generally directed and anisotropic. These are often classified based on their symmetry with respect to a molecular plane as **Sigma** (σ) bonds and **Pi** (π) bonds. In the general case, atoms form bonds that are intermediate between ionic and covalent, depending on the relative electronegativity of the atoms involved. Bonds of this type are known as polar covalent bonds.

1.5 TYPES AND PROPERTIES OF CHEMICAL BOND

The strength of chemical bonds varies considerably; there are ‘Strong Bonds’ or Primary Bonds’, such as covalent, ionic and metallic bonds, and ‘Weak Bonds’ or ‘Secondary Bonds’, such as dipole–dipole interactions, the London dispersion force and hydrogen bonding.

Strong Bonds or Primary Bonds: Strong chemical bonds are the intramolecular forces that hold atoms together in molecules. A strong chemical bond is formed from the transfer or sharing of electrons between atomic centres and relies on the electrostatic attraction between the protons in nuclei and the electrons in the orbitals. The types of strong bond differ due to the difference in

NOTES

electronegativity of the constituent elements. A large difference in electronegativity leads to more polar (ionic) character in the bond. There are following types of strong chemical bonds:

- Covalent Bonds
- Ionic Bonds
- Metallic Bonds

Covalent Bonds: Covalent bonding is a common type of bonding in which two or more atoms share valence electrons more or less equally. The simplest and most common type is a single bond in which two atoms share two electrons. Other types include the double bond, the triple bond, one- and three-electron bonds, the three-centre two-electron bond and three-centre four-electron bond.

In non-polar covalent bonds, the *electronegativity* difference between the bonded atoms is small, typically 0 to 0.3. Bonds within most organic compounds are described as covalent. Figure 1.2 shows Methane (CH_4), in which each hydrogen forms a covalent bond with the carbon.

Molecules that are formed primarily from non-polar covalent bonds are often immiscible in water or other polar solvents, but much more soluble in non-polar solvents, such as hexane.

A polar covalent bond is a covalent bond with a significant ionic character. This means that the two shared electrons are closer to one of the atoms than the other, creating an imbalance of charge. Such bonds occur between two atoms with moderately different electronegativity's and give rise to *dipole–dipole interactions*. The electronegativity difference between the two atoms in these bonds is 0.3 to 1.7. Figure 1.2 illustrates the non-polar covalent bonds in Methane (CH_4). The Lewis structure shows electrons shared between C and H atoms.

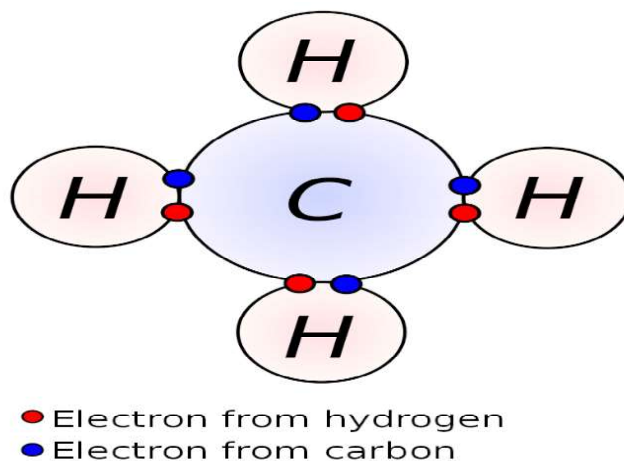


Fig.1.2 *Non-Polar Covalent Bonds in Methane (CH_4)*

Single and Multiple Bonds

A single bond between two atoms corresponds to the sharing of one pair of electrons. The Hydrogen (H) atom has one valence electron. Two Hydrogen atoms can then form a molecule, held together by the shared pair of electrons. Each H atom now has the noble gas electron configuration of Helium (He). The pair of shared electrons forms a single covalent bond. The electron density of these two bonding electrons in the region between the two atoms increases from the density of two non-interacting H atoms.

A double bond has two shared pairs of electrons, one in a sigma bond and one in a pi bond with electron density concentrated on two opposite sides of the internuclear axis. A triple bond consists of three shared electron pairs, forming one sigma and two pi bonds, as in the case of nitrogen. Quadruple and higher bonds are very rare and occur only between certain transition metal atoms. Figure 1.3 illustrates two p-orbitals forming a pi-bond.

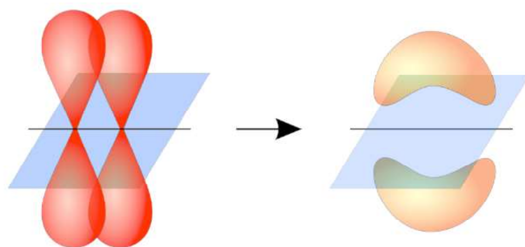


Fig.1.3 *Two p-Orbitals forming a pi-Bond*

Coordinate Covalent Bond (Dipolar Bond)

A coordinate covalent bond is a covalent bond in which the two shared bonding electrons are from the same one of the atoms involved in the bond. For example, Boron Trifluoride (BF_3) and Ammonia (NH_3) form an adduct or coordination complex F_3B !NH_3 with a B–N bond in which a lone pair of electrons on N is shared with an empty atomic orbital on B. BF_3 with an empty orbital is described as an electron pair acceptor or Lewis acid, while NH_3 with a lone pair that can be shared is described as an electron-pair donor or Lewis base. The electrons are shared roughly equally between the atoms in contrast to ionic bonding. Such bonding is shown by an arrow pointing to the Lewis acid.

Transition metal complexes are generally bound by coordinate covalent bonds. For example, the ion Ag^+ reacts as a Lewis acid with two molecules of the Lewis base NH_3 to form the complex ion $Ag(NH_3)_2^+$, which has two Ag ! N coordinate covalent bonds. Figure 1.4 shows the adduct of 'Ammonia' and 'Boron Trifluoride'. An adduct (from the Latin adductus, 'Drawn Toward' alternatively, a

contraction of 'Addition Product') is a product of a direct addition of two or more distinct molecules, resulting in a single reaction product containing all atoms of all components. The resultant is considered a distinct molecular species.

NOTES

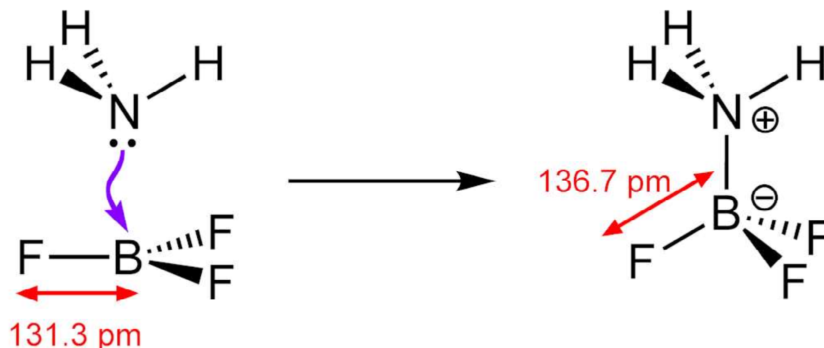


Fig.1.4 *Adduct of Ammonia and Boron Trifluoride*

Ionic Bond: Ionic bonding is a type of electrostatic interaction between atoms that have a large electronegativity difference. There is no precise value that distinguishes ionic from covalent bonding, but an electronegativity difference of over 1.7 is likely to be ionic while a difference of less than 1.7 is likely to be covalent. Ionic bonding leads to separate positive and negative ions. Ionic charges are commonly between –3eto +3e. Ionic bonding commonly occurs in metal salts, such as Sodium Chloride (salt). A typical feature of ionic bonds is that the species form into ionic crystals, in which no ion is specifically paired with any single other ion in a specific directional bond. Rather, each species of ion is surrounded by ions of the opposite charge, and the spacing between it and each of the oppositely charged ions near it is the same for all surrounding atoms of the same type. It is thus no longer possible to associate an ion with any specific other single ionised atom near it. This is a situation unlike that in covalent crystals, where covalent bonds between specific atoms are still discernible from the shorter distances between them, as measured via such techniques as X-ray diffraction.

Ionic crystals may contain a mixture of covalent and ionic species, as for example salts of complex acids, such as Sodium Cyanide (NaCN). X-ray diffraction shows that in NaCN, for example the bonds between Sodium Cations (Na^+) and the Cyanide Anions ($CN^–$) are ionic, with no sodium ion associated with any particular cyanide. However, the bonds between C and N atoms in cyanide are of the covalent type, so that each carbon is strongly bound to just one nitrogen, to which it is physically much closer than it is to other carbons or nitrogen's in a sodium cyanide crystal.

When such crystals are melted into liquids, the ionic bonds are broken first because they are non-directional and allow the charged species to move freely. Similarly, when such salts dissolve into water, the ionic bonds are typically broken by the interaction with water but the covalent bonds continue to hold. For example, in solution, the cyanide ions, still bound together as single $CN^–$ ions, move

independently through the solution, as do Sodium Ions, as Na^+. In water, charged ions move apart because each of them are more strongly attracted to a number of water molecules than to each other. The attraction between ions and water molecules in such solutions is due to a type of weak dipole-dipole type chemical bond. In melted ionic compounds, the ions continue to be attracted to each other, but not in any ordered or crystalline way.

Metallic Bonding: Metallic bonding is a type of chemical bonding that arises from the electrostatic attractive force between conduction electrons (in the form of an electron cloud of delocalised electrons) and positively charged metal ions. It may be described as the sharing of free electrons among a structure of positively charged ions (cations). Metallic bonding accounts for many physical properties of metals, such as strength, ductility, thermal and electrical resistivity and conductivity, opacity, and lustre.

Metallic bonding is not the only type of chemical bonding a metal can exhibit, even as a pure substance. For example, elemental Gallium consists of covalently-bound pairs of atoms in both liquid and solid-state these pairs form a crystal structure with metallic bonding between them. Another example of a metal–metal covalent bond is the mercurous ion (Hg_2^{2+}).

In metallic bonding, bonding electrons are delocalised over a lattice of atoms. By contrast, in ionic compounds, the locations of the binding electrons and their charges are static. The free movement or delocalisation of bonding electrons leads to classical metallic properties, such as lustre (surface light reflectivity), electrical and thermal conductivity, ductility, and high tensile strength.

Weak Bonds or Secondary Bonds: Secondary bonds are bonds of a different kind to the primary ones. They are weaker in nature and are broadly classified as Van der Waal's forces and hydrogen bonds. These bonds are due to atomic or molecular dipoles, both permanent and temporary. For example, water molecule is made of one oxygen and two hydrogen atoms. There are following types of weak chemical bonds:

- Dipole–Dipole Interactions
- London Dispersion Force
- Hydrogen Bonding

Dipole–Dipole Interactions

Regular Dipole: Dipole–dipole interactions are electrostatic interactions between molecules which have permanent dipoles. This interaction is stronger than the London forces but is weaker than ion-ion interaction because only partial charges are involved. These interactions tend to align the molecules to increase attraction (reducing potential energy). An example of a dipole–dipole interaction can be seen in Hydrogen Chloride (HCl): the positive end of a polar molecule will attract the negative end of the other molecule and influence its position. Polar molecules have a net attraction between them. Examples of polar molecules include Hydrogen

NOTES

Chloride (HCl) and Chloroform ($CHCl_3$). Often molecules contain dipolar groups of atoms, but have no overall dipole moment on the molecule as a whole. This occurs if there is symmetry within the molecule that causes the dipoles to cancel each other out. This occurs in molecules, such as tetra chloromethane and carbon dioxide. The dipole–dipole interaction between two individual atoms is usually zero, since atoms rarely carry a permanent dipole.

Ion–Dipole and Ion–Induced Dipole Forces: Ion–dipole and ion–induced dipole forces are similar to dipole–dipole and dipole–induced dipole interactions but involve ions, instead of only polar and non-polar molecules. Ion–dipole and ion–induced dipole forces are stronger than dipole–dipole interactions because the charge of any ion is much greater than the charge of a dipole moment. Ion–dipole bonding is stronger than hydrogen bonding. An ion–dipole force consists of an ion and a polar molecule interacting. They align so that the positive and negative groups are next to one another, allowing maximum attraction. An important example of this interaction is hydration of ions in water which give rise to hydration enthalpy. The polar water molecules surround themselves around ions in water and the energy released during the process is known as hydration enthalpy. The interaction has its immense importance in justifying the stability of various ions (like Cu^{2+}) in water. An ion–induced dipole force consists of an ion and a non-polar molecule interacting. Like a dipole–induced dipole force, the charge of the ion causes distortion of the electron cloud on the non-polar molecule.

London Dispersion Force: London Dispersion Forces (LDF), also known as dispersion forces, London forces, instantaneous dipole–induced dipole forces, Fluctuating Induced Dipole Bonds or loosely as Van der Waals forces, are a type of force acting between atoms and molecules that are normally electrically symmetric, that is, the electrons are symmetrically distributed with respect to the nucleus. They are part of the Van der Waals forces. The LDF is named after the German physicist Fritz London.

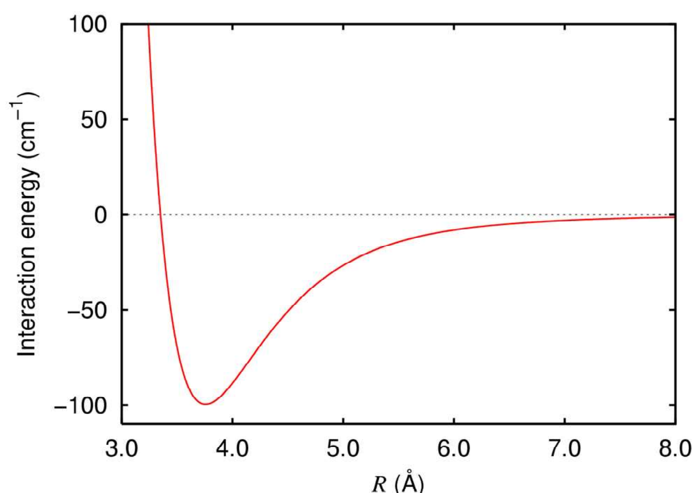


Fig.1.5 *Interaction Energy of an Argon Dimer*

NOTES

Figure 1.5 illustrates the interaction energy of an Argon Dimer. The long-range segment is because of the London Dispersion Forces (LDFs).

Hydrogen Bonding: A hydrogen bond (H-bond) is a primarily electrostatic force of attraction between a Hydrogen (H) atom which is covalently bound to a more electronegative atom or group, and another electronegative atom bearing a lone pair of electrons the hydrogen bond Acceptor (Ac). Such an interacting system is generally denoted as Dn — H ······ Ac, where the solid line denotes a polar covalent bond, and the dotted or dashed line indicates the hydrogen bond. Hydrogen bonds can be intermolecular (occurring between separate molecules) or intramolecular (occurring among parts of the same molecule). Depending on the nature of the donor and acceptor atoms which constitute the bond, their geometry, and environment, the energy of a hydrogen bond can vary between 1 and 40 kcal/mol. This makes them somewhat stronger than a Van der Waals interaction, and weaker than fully covalent or ionic bonds. This type of bond can occur in inorganic molecules, such as water and in organic molecules, such as DNA and proteins.

The hydrogen bond is responsible for many of the anomalous physical and chemical properties of compounds of N, O, and F. In particular, intermolecular hydrogen bonding is responsible for the high boiling point of water (100 °C) compared to the other group 16 hydrides that have much weaker hydrogen bonds. Intramolecular hydrogen bonding is partly responsible for the secondary and tertiary structures of proteins and nucleic acids. It also plays an important role in the structure of polymers, both synthetic and natural.

Figure 1.6 illustrates the model of Hydrogen Bonds (1) between the molecules of water.

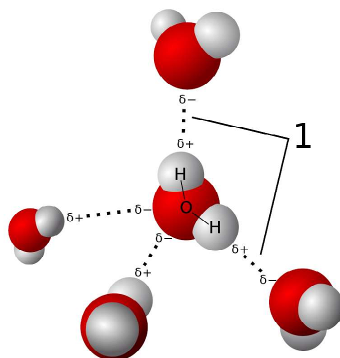


Fig. 1.6 *Model of Hydrogen Bonds (1) between Molecules of Water*

A hydrogen atom attached to a relatively electronegative atom is the hydrogen bond donor. C–H bonds only participate in hydrogen bonding when the carbon atom is bound to electronegative substituents, as is the case in Chloroform, $CHCl_3$. In a hydrogen bond, the electronegative atom not covalently attached to the hydrogen is named proton acceptor, whereas the one covalently bound to the hydrogen is named the proton donor. While this nomenclature is recommended by the

NOTES

International Union of Pure and Applied Chemistry (IUPAC), it can be misleading, since in other donor-acceptor bonds, the donor/acceptor assignment is based on the source of the electron pair. In the hydrogen bond donor, the H centre is protic. The donor is a Lewis base. Hydrogen bonds are represented as H···Y system, where the dots represent the hydrogen bond. Liquids that display hydrogen bonding (such as, water) are called associated liquids. The hydrogen bond is often described as an electrostatic dipole-dipole interaction. However, it also has some features of covalent bonding: it is directional and strong, produces interatomic distances shorter than the sum of the Van der Waals radii, and usually involves a limited number of interaction partners, which can be interpreted as a type of valence. These covalent features are more substantial when acceptors bind hydrogens from more electronegative donors.

1.6 POLYMERISATION OF ORGANIC MOLECULES

Polymers are one of the most important products, which find an important place in every walk of modern civilisation. The term polymer (Greak word: *poly* + *meros*, means, many parts) denotes a molecule produced by the repetition of some simpler unit, called the mer or the monomer.

The term *macromolecule* (big molecule) is also often used to cover the large molecule of complex structure. A naturally occurring macromolecule is insulin, a protein hormone, which occurs in the pancreas, and is best known agent to lower blood sugar in diabetic patients. It has the repeating units with amide linkages,

i.e., $(\text{—}\overset{\overset{\displaystyle R}{|}}{C}H\text{—}\overset{\overset{\displaystyle O}{||}}{C}\text{—NH—})_n$, where $n = 51$ and R has sixteen variations. The science of macromolecules is divided between biological and non-biological materials, each having vital importance in our daily life. Biological polymers, i.e., proteins, nucleic acids (DNA, RNA), starch, cellulose and enzymes are complex macro molecules which form the very foundation of life and intelligence and provide much of the food for the existance of man. This chapter however, is concerned mainly with the chemistry of some polymers. These are primarily the synthetic materials used for plastics, fibres, and elastomers and a few naturally occurring polymers, such as rubber, wool and cellulose. The polymers are truly indispensible to mankind, as they are used to meet the basic needs-clothing, shelter, communication, and transportation, as well as to the conveniences of modern living.

The name of a polymer is usually derived from the name of the monomer (repeat unit) by prefixing the word poly to it. To illustrate the polymerisation product of ethylene is known as polyethylene and that of sytrene is called polystyrene. The number of repeating units in the chain is called the degree of polymerisation (DP) and specifies the length of the chain. Degree of polymerisation is denoted by the

NOTES

letter n or P. The molecular weight of the polymer is the product of the molecular weight of the repeat unit and the degree of polymerisation, i.e.,

$$M_{\text{poly}} = n \times M_m$$

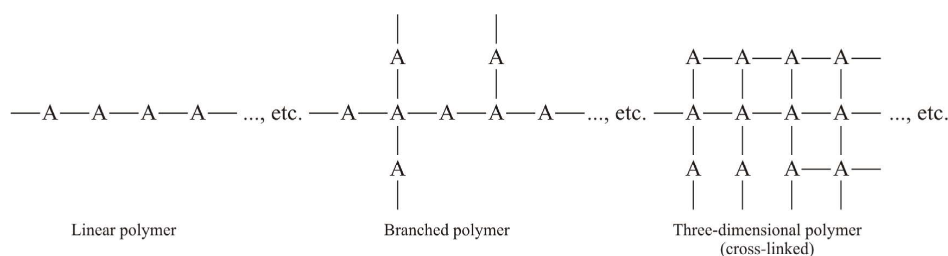
or

$$n \text{ (Degree of polymerisation)} = \frac{M_{\text{poly}} \text{ (mol. wt. of polymer)}}{M_m \text{ (mol. wt. of repeat unit)}}$$

The degree of polymerisation may vary over a wide range, i.e., from a few units to 10,000 and even more. Polymers with high P are called high polymers and are mostly useful for plastics, rubbers or fibres, etc., while those with low P are known as oligomers.

Classification of Polymers

A polymer may consist of monomers of identical or of different chemical structure. If it has identical units then it is known as homopolymer, whereas a polymer containing several types of monomeric units in its chain is known as copolymer, or mixed polymer. In some cases the repetition is linear and a chain is built up from its links. However, in some cases the chains are branched or interconnected, to form three dimensional structures.



Copolymers may also be linear, branched or three-dimensional. The monomer residues, in co-polymer molecules may be arranged in the chain regularly or at random, according to the law of chance. Copolymers of the former group are called regular-copolymers and those of latter type, statistical or irregular copolymers.

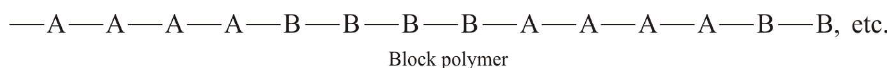
Copolymers with long sequence of two monomers can have two arrangements of long chains:

1. linear copolymers in which the units of each type form fairly continuous sequence (blocks) is known as block copolymers.
2. Branched copolymer, with monomer of one kind grafted into a backbone of the second monomer type is termed as graft copolymer.

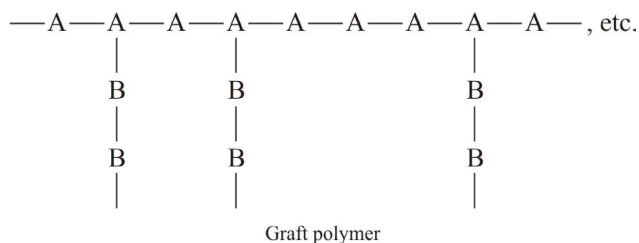
According to the structure of main chain, all polymers are classified into homochain and heterochain polymers. Homochain polymers contain the chains composed of atoms of the same element, e.g., carbon, sulphur phosphorus etc.

NOTES

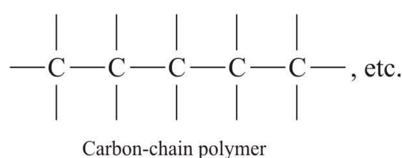
In heterochain polymers, the main chain is made up of atoms of different species, e.g.,



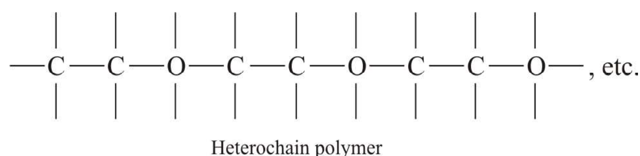
Block polymer



Graft polymer



Carbon-chain polymer



Heterochain polymer

On the basis of chemical composition, the polymers can also be classified as:

1. **Organic polymers:** Organic polymers include compounds containing carbon hydrogen, oxygen nitrogen, sulphur and halogen atoms. Oxygen, nitrogen or sulphur may also be present in the backbone chain.
2. **Elemento-organic (hetero-organic) polymers:** These include the following classes:
   (*i*) Compounds containing carbon atoms and hetero atoms (except for nitrogen, sulphur and oxygen atoms) in their chains.
   (*ii*) Compounds with inorganic chains if they contain side groups with carbon atoms connected directly to the chain.
   (*iii*) Compounds having carbon atoms in the main chain and hetero-atoms (except for nitrogen, sulphur, oxygen and halogen atoms) in side groups connected directly to the carbon atoms of the chain. For example, polysiloxanes, polytitanoxanes, etc.
3. **Inorganic polymers:** These include polymers containing no carbon atoms. Their are composed of different atoms joined by chemical bonds, while weaker intermolecular forces act between their chains. Polysilanes, polygermanes, polysilicic acid, polyphosphates, polyarsenates, etc., are examples of inorganic polymers.

Polymerisation Processes

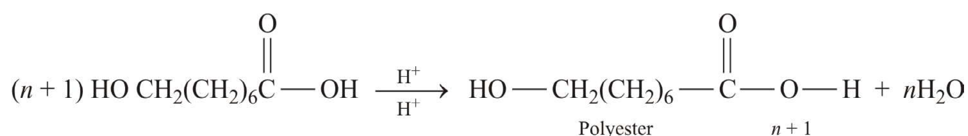
The processes of polymerisation were classified by the W H Carothers (1926) into two groups, i.e., *condensation* and *addition* polymerisations. In 1953, Flory amended Carother's original distinction between addition and condensation polymerisation. He laid special emphasis on the mechanisms by which the two types of polymerisation proceed. It was observed that condensation polymerisation was preceded by the stepwise intermolecular condensation of reactive groups and the addition polymerisation resulted from chain reactions involving some sort of active centres.

NOTES

Thus, the two classes of polymerzation are:

Condensation Polymerisation

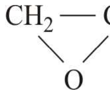
This is entirely analogous to condensation in low molecular weight compounds. For the formation of condensation polymer there is union between two polyfunctional molecules to produce the large polyfunctional molecule. The process involves the elimination of a small molecule, such as water, ammonia, etc. The reaction continues until almost all of the reagents is used up. Since there is an equilibrium between reactants and products, the rate of conversion can be controlled by the rate of removal of one of the products. A good example is, esterification, where water is eliminated between an acid and an alcohol.



Here, the rate of conversion can be controlled by the rate of removal of water.

Addition or Chain-reaction Polymerisation

This type of polymerisation involves chain reactions. The chain carrier in such reactions may be an ion or a reactive substance with one unpaired electron called a free radical. Chain polymerisation is characteristic of compounds with multiple bonds, e.g., ethylene $CH_2{=}CH_2$. isobutylene $(CH_3)_2\ C{=}CH_2$ and vinyl chloride $CH_2{=}CHCl$, or of unstable rings containing heteroatoms, e.g., ethylene oxide



. Depending upon the active centre, which may be a free radical or an ion, the reaction is one of radical or ionic polymerisation, respectively. A free radical is produced by the decomposition of an initiator. The free radical then attacks to open the double bond of a vinyl monomer or ring or a cyclic compound and adds to it, with an electron remaining unpaired. Within a very short span of time (usually a few seconds or less) many more monomers add successively to form a long chain with active centers. Finally, chain termination results from saturation of the macro radical and may occur by the combination of free radicals, disproportionation of chain transfer.

NOTES

With few exceptions, chain-reaction polymerisation results in the formation of homochain polymers, where as step-reaction polymerisation produces heterochain polymers. Polymerisations are classified without regard to loss of a small molecule (e.g., polyurethanes are formed by step-reaction polymerisation or type of interunit linkage (e.g., phenol-formaldehyde resins results from stepwise polymerisation even though they lack interunit functional groups). In case the differentiation is required on the basis of mechanism, the terms *step-reaction* and *chain-reaction* are used; but to avoid confusion, the common terms *condensation* and *addition* are permissible.

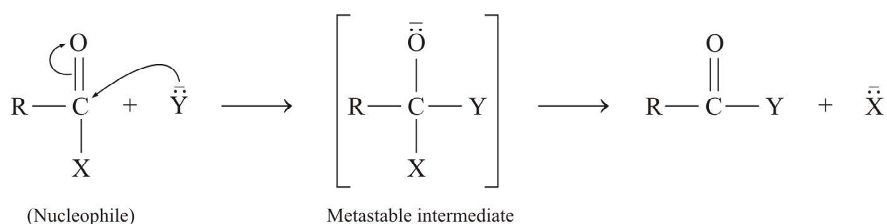
Chemistry of Polymerisation

The reactions by which complex macromolecules are formed are not completely understood. The process of building up polymers from simple repeating units (monomers) can proceed with many variations. As noted earlier, on the basis of mechanisms the two types of polymers are formed. Condensation polymers are usually formed by the step-wise intermolecular condensation of the reactive groups, addition polymers ordinarily result from chain reactions involving some sort of active centres. The chemistry of each type is discussed in this section except those which are not yet confirmed in detail.

Stepwise Polymerisation

Most of the stepwise polymerisations are, stoichometrically, condensation polymerisation. A monomer can be converted into polymer by any reaction that creates new bonds. Carothers defined this number of new-bonds as the functionality of a monomer in a given reaction. As the number of bonds formed depends upon the number of reactive functional groups, the *functionality* of a monomer can also be defined as the average number of reactive functional group per molecule. In a condensation reaction, the type of the product formed is determined by the functionality of the monomer. It should be obvious that a monofunctional monomer gives only a low-molecular-weight product. A functionality too can lead to linear structure. Polyfunctional monomer, with more than two functional groups per molecule, give branched or cross-linked (three-dimensional) polymers. The linear and the three-dimensional polymers differ widely in their properties.

The most important reaction which has been used for the preparation of condensation polymers is the addition and elimination at the carbonyl double bond of carboxylic acids and their derivatives. The generalized reaction is

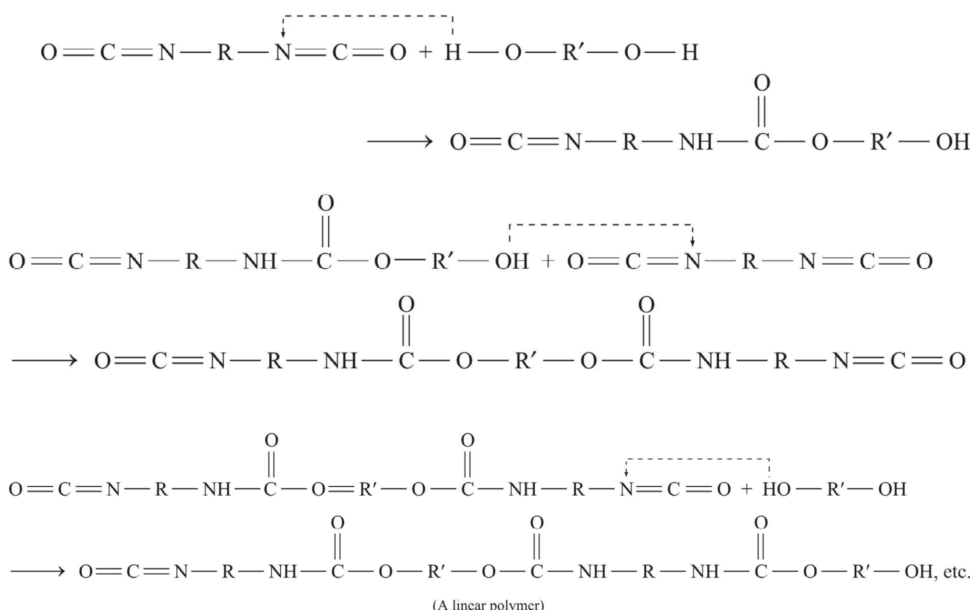


NOTES

where R may be alkyl or aryl group, *X* may be OH, OR′, NH_2. NHR, $\text{O}-\overset{\text{O}}{\overset{\|}{\text{C}}}-\text{R}$, or Cl; and Y may be R′O–, R′OH, $R'NH_2$ or $R'COO^-$. The metastable intermediate can either return to the original state by Y or proceed to the final state by eliminating X. The formation of polymers, polyamides, nylon—66 poly (ethylene terephathalate), polyurethanes, polyureas, polysulphonamides, polyphenyl esters, etc. provides some typical examples of this reaction.

In simple terms, it can be stated that stepwise polymerisation is the combination of several molecules by stepwise addition of the monomer molecules to one another as a result of migration of some mobile atom, usually a hydrogen atom, from one molecule to another. The formation of polyurethanes from di-isocyanates and dihydric alcohols is illustrated below:

A trihydric alcohol will give a cross-linked polymer.



(A linear polymer)

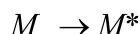
Chain Polymerisation

Chain polymerisation is characteristic of unsaturated monomers. Ethenic polymerisation is an economically important class whose kinetics typify chain polymerisation. The terms 'vinyl', 'olefin' or 'addition' polymerisation are often used, although they are more restrictive. Flory showed conclusively that chain polymerisation proceeds by and requires the steps of interaction, propagation, and termination typical of chain reactions in low-molecular-weight species. The

NOTES

three stages which are essential to the formation of a high polymer can be represented as:

1. **Initiation:** This involves the creation of an 'active' centre, and can also be termed as activation of the monomer molecule, i.e.,



$$M \rightarrow M^*$$

(Excited monomer molecule or active centre)

2. **Propagation:** Propagation involves the addition of more monomer species to the chain end. This occurs usually very rapidly (mol. wt. 10^7 in one-tenth of a second) to final molecular weight value, as shown below:



$$M^* + M \rightarrow M_1^*$$



$$M_1^* + M \rightarrow M_2^*$$



$$M_2^* + M \rightarrow M_3^*$$

....
....
....



$$M_{n-1}^* + M \rightarrow M_n^* \text{ (Activated growing polymer molecule)}$$

3. **Termination:** In this step there is disappearance of the 'active' centre



$$M_n^* \rightarrow M_n \text{ (Inactivated polymer molecule)}$$

The active centre in chain polymerisation reactions may be a free radical or an ion, and the reaction is one of radical or ionic polymerisation, respectively.

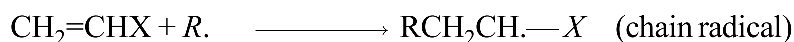
Radical Polymerisation

The intermediates having an odd number of electrons (an unpaired electron) are known as free radicals. They can be generated in a number of ways, including thermal decomposition of peroxides or hydroperoxides; photolytic decomposition of covalently bonded compounds; dissociation of covalent bonds by high energy radiation (α-, β-, γ-rays); oxidation reduction reactions; and electrochemical interaction. However, the stability of the free radicals varies widely. Tertiary radicals are more stable and less reactive than secondary radicals, which are in turn more stable and less reactive than the primary ones. The benzyl radical is less reactive than phenyl radical, the allyl is quite stable and is quite unreactive.

1. **Initiation:** Once the primary free radicals (free radicals produced in the first stage) are produced by physical or chemical effects in the presence of a vinyl monomer (CH_2=CHX), the radical adds to the ethylenic double bond of an unexcited monomer molecule with the regeneration of another radical. Let the radical formed by decomposition of initiator *R–R* is designated *R*



$$R - R \xrightarrow{\text{decomposition}} 2R \quad \text{(Primary free radical)}$$



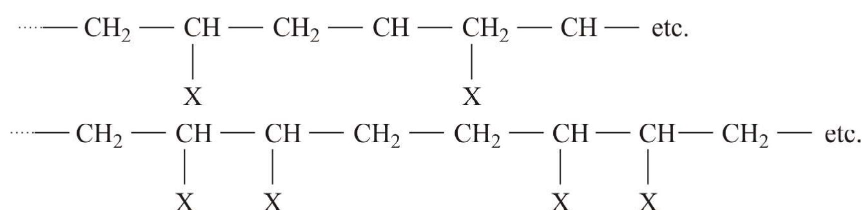
$$CH_2{=}CHX + R. \longrightarrow RCH_2CH.{—}X \quad \text{(chain radical)}$$

The regeneration of the new radical is characteristic of chain reactions as this is capable of further interaction with the initial monomers. The efficiency

NOTES

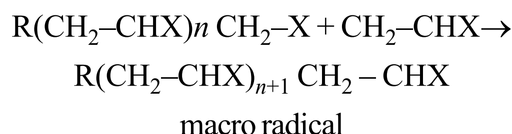
with which radicals initiate can be estimated by comparing the amount of initiator decomposed with the number of polymer chains formed. Most initiator in vinyl polymerisations have 60–100 per cent efficiency. The low efficiency is mainly due to the recombination of the radical pairs before they move apart, this is known as the *cage-effect*.

2. **Propagation:** The chain radical formed in the initiation step is capable of adding successive monomers to from macroradicals. The chain propagation reaction determines the rate of the polymerisation, the molecular mass of polymer, the structure of the polymer chain, i.e., the mode of monomer addition ('head-to-head' or head-to-tail)*, the degree of branching, etc.



Addition according to the former scheme is called head-to-tail, configuration in which the substituents occur on alternate carbon, atoms, and that according to the latter scheme head-to-head or tail-to-tail In most polymerisations, monomers combine according to most favoured (steric factors) head-to-tail scheme.

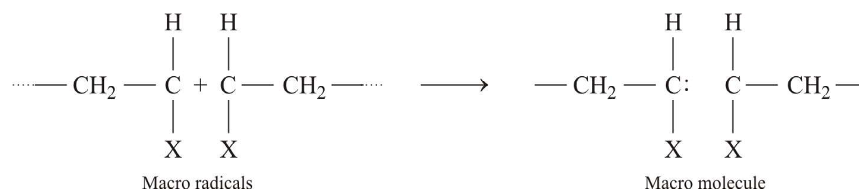
The propagation step is given by following reaction:



$$R(CH_2–CHX)n\ CH_2–X + CH_2–CHX \rightarrow$$
$$R(CH_2–CHX)_{n+1}\ CH_2 – CHX$$
macro radical

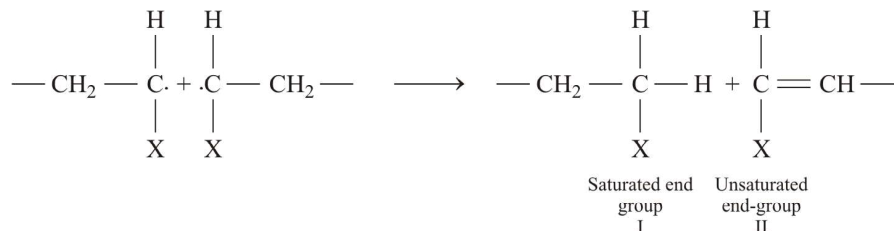
3. **Termination:** Propagation would continue until the supply of monomer is exhausted. The radicals also have the strong tendency to react in pairs to form a paired-electron convalent bond causing thereby loss of activity. In fee radical polymerzation, this tendency is compensated by the small concentration of radical species compared to monomers. Chain termination results from saturation (deactivation) of the macroradical and may occur in two ways:

   (*i*) **Combination or coupling:** Combination consists in the neutral saturation of two macroradicals or of a macroradical and a low molecular weight free radical:



NOTES

(*ii*) **Disproportionation:** This involves the transfer of a hydrogen atom from one macroradical to another to form two macromolecules with one saturated and one unsaturated end group.



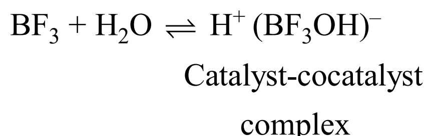
Chemical substances capable of reacting with free radicals to terminate the reaction chain are called polymerisation *inhibitors*, for example–hydroquinone and trinitrobenzene. Chemical compounds which are only chain transfer agents and do not affect the rate of polymerisation but determine the molecular mass of polymer are termed as polymerisation regulators or modifiers.

Ionic Polymerisation

Chain reaction polymerisation can occur by several mechanisms other than those involving free radicals as discussed earlier. The most important among these are the reactions in which the chain carriers are carbonium ions (cationic polymerisation) or carbanions (anionic polymerisation). Ionic polymerisation proceeds in the presence of catalysts and is also called catalytic polymerisation. The reactivity of the ethenic monomers to polymerisation by radicals, ions and complexing agents varies with the structure in a manner which can be correlated though not always quantitatively predicated. It can be seen that, for the vinyl monomer (CH_2=CH–X), cationic initiation is favoured when *X* is electron releasing and anionic when X is electron-withdrawing. This is because monomers with electron releasing groups attached to the double-bonded carbons are capable of forming stable carbonium ions, where as monomers with electron with-drawing substituents form stable carbanions. Ionic polymerisations tend to be very rapid even at low temperatures. The polymerisation of isobutylene with $AlCl_3$ or BF_3 is carried out commercially at —100ºC and an estimate of the life-time of a growing chain of isobutylene in this case is about 10^{-6} sec, much shorter than the usual life time of a free-radical chain.

1. **Cationic or carbonium polymerisation:** This involves the formation of a carbonium ion which is a polar compound with tricovalent carbon atom carrying a positive charge, R–CH–R. Typical catalysts for cationic polymerisation are compounds with pronounced electron acceptor properties (Lewis acids), e.g., $AlCl_3$, $AlBr_3$, $SnCl_4$, BF_3, H_2SO_4 and other strong acids. The cationic polymerisation involves the carbonium ion as the chain carrier. Carbonium ion ineracts with a monomer molecule, the reaction of chain growth being accompanied with the communication of positive charge along the chain. Consequently, the growing chain itself is a cation with a

molecular mass increasing in the course of polymerisation. For example, the polymerisation of isobutylene in the presence of boron trifluoride catalyst can be represented as:

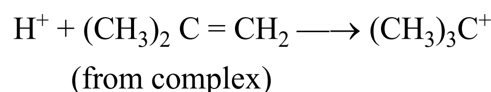


$$BF_3 + H_2O \rightleftharpoons H^+ (BF_3OH)^-$$

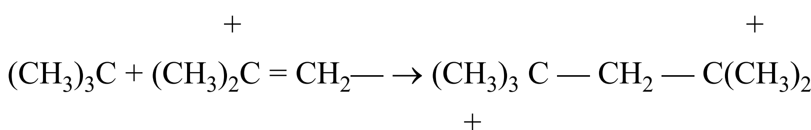
Catalyst-cocatalyst complex

NOTES

The catalyst-cocatalyst complex formed in step (I) donates a proton to an isobutylene molecule to give carbonium ion.



$$\underset{\text{(from complex)}}{H^+} + (CH_3)_2 C = CH_2 \longrightarrow (CH_3)_3C^+$$



$$(CH_3)_3\overset{+}{C} + (CH_3)_2C = CH_2 \longrightarrow (CH_3)_3 C - CH_2 - \overset{+}{C}(CH_3)_2$$



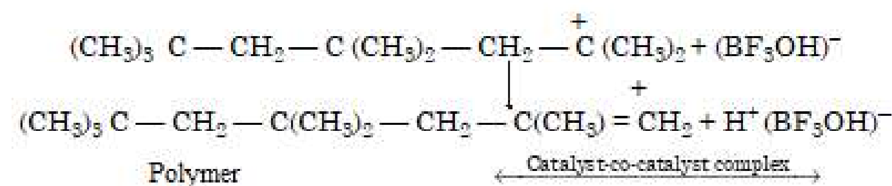
$$(CH_3)_3 C - CH_2 - \overset{+}{C}(CH_3)_2 + (CH_3)_2 C = CH_2 \longrightarrow$$



$$(CH_3)_3 C - CH_2C (CH_3)_2 - CH_2 - \overset{+}{C} (CH_3)_2$$

Thus, there is reformation of a carbonium ion at the end of each step by the 'head-to-tail' addition of monomer to ion.

Chain termination apparently occurs as a result of the mutual collision of the ends of growing ion to yield a polymer molecule with terminal unsaturation and the original complex



$$(CH_3)_3 C - CH_2 - C(CH_3)_2 - CH_2 - \overset{+}{C}(CH_3)_2 + (BF_3OH)^-$$

$$\downarrow$$

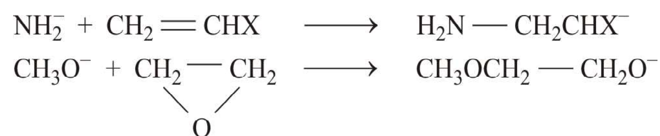
$$\underbrace{(CH_3)_3 C - CH_2 - C(CH_3)_2 - CH_2 - C(CH_3) = CH_2}_{\text{Polymer}} + \underbrace{H^+ (BF_3OH)^-}_{\text{Catalyst-co-catalyst complex}}$$

The third component present in low concentration which usually has pronounced effects on polymerisation is called cocatalyst. The efficiency of the catalyst is dependent on the acid strength of the catalyst-cocatalyst complex. In cationic polymerisation, the catalyst is not in the macromolecule.

2. **Anionic or carbanion polymerisation:** Anionic polymerisation involves the formation of a carbanion, a compound with a trivalent carbon atom carrying a negative charge. Catalyst for carbanion polymerisation include, alkali metals, alkali metal amides, alkoxides, alkyls, aryls, hydroxides and cyanides. They are electron releasing groups. Polymerisation occurs by the carbanion mechanism in the case of monomers containing electronegative substituents at one of the carbon atoms connected by a double bond. The

NOTES

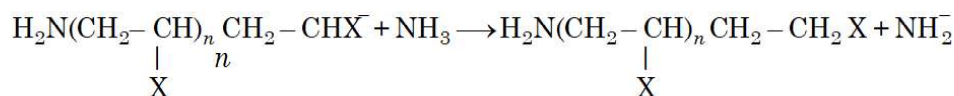
initiation of ionic chains involves the addition of a negative ion to the monomers, with the opening of a ring or bond and growth at one end



The more basic the ion (anion), the better it serves to initiate chains. Thus, $C_6H_5CH_2$- is powerful initiator than $\propto H_2^-$ which is stronger than OH^- in the anionic polymerisation of styrene.

The chain growth reaction is always accompanied with a transfer of a negative charge along the chain, consequently, the growing chain is always an anion of growing size.

Chain termination occurs as a result of the collision of a growing ion with a molecule of the medium, such as an ammonia molecule:



Thus, energy growing chain contains the $-NH_2$ group. The termination step is always unimolecular and usually be transfer.

It anionic polymerisation, the end group of a growing molecule possesses high activity and great stability. Hence, the polymers obtained by anionic polymerisation method retain active centres at the end of the chain, which are capable of initiation the polymerisation of monomers. Such polymers are called living polymers. The polymer can be 'killed' by addition of a terminating agent, like water at the end of the reaction. Anionic polymerisation has great advantages, since spontaneous chain termination does not occur.

Check Your Progress

1. Explain about the biophysics.
2. Elaborate on the structure of an atoms.
3. Analyse the structure of molecules.
4. What do you understand by chemical bond?
5. What is strong chemical bond?
6. Analyse the polymerisation process.
7. Explain the chemistry of polymerisation.
8. Define the chain polymerisation.

1.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. Biophysics is the field that applies the theories and methods of physics to understand how biological systems work. Biophysics has been critical to understanding the mechanics of how the molecules of life are made, how different parts of a cell move and function, and how complex systems in our bodies—the brain, circulation, immune system, and others— work. Biophysics is, therefore, an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena. Biophysics covers all scales of biological organisation, from molecular to organismic and populations.
2. Typically, an atom is the smallest unit of matter that retains all of the chemical properties of an element. Atoms combine to form molecules, which then interact to form solids, gases, or liquids. For example, water is composed of hydrogen and oxygen atoms that have combined to form water molecules. Several biological processes are dedicated to breaking down molecules into their component atoms so that they can be reconstructed into a more useful molecule. Atoms are made up of particles called protons, neutrons, and electrons, which are responsible for the mass and charge of atoms.
3. A molecule is an electrically neutral group of two or more atoms held together by chemical bonds. Molecules are distinguished from ions by their lack of electrical charge. In quantum physics, organic chemistry, and biochemistry, the distinction from ions is dropped and molecule is often used when referring to polyatomic ions. According to the kinetic theory of gases, the term molecule is often used for any gaseous particle regardless of its composition. This violates the definition that a molecule contain two or more atoms, since the noble gases are individual atoms.
4. Chemical bonds are forces that hold the atoms together in a molecule. They are a result of strong intramolecular interactions among the atoms of a molecule. The valance (outermost) electrons of the atoms participate in chemical bonds. When two atoms approach each other, these outer electrons start to interact. Although electrons repel each other, they are attracted to the protons within atoms. The interplay of forces results in the formation of bonds between the atoms.
5. Strong chemical bonds are the intramolecular forces that hold atoms together in molecules. A strong chemical bond is formed from the transfer or sharing of electrons between atomic centres and relies on the electrostatic attraction between the protons in nuclei and the electrons in the orbitals. The types of strong bond differ due to the difference in electronegativity of the constituent elements. A large difference in electronegativity leads to more polar (ionic) character in the bond.

NOTES

6. The processes of polymerisation were classified by the W. H. Carothers (1926) into two groups, i.e., condensation and addition polymerisations. In 1953, Flory amended Carothers original distinction between addition and condensation polymerisation. He laid special emphasis on the mechanisms by which the two types of polymerisation proceed. It was observed that condensation polymerisation was preceded by the stepwise intermolecular condensation of reactive groups and the addition polymerisation resulted from chain reactions involving some sort of active centres.
7. The reactions by which complex macromolecules are formed are not completely understood. The process of building up polymers from simple repeating units (monomers) can proceed with many variations. On the basis of mechanisms the two types of polymers are formed. Condensation polymers are usually formed by the step-wise intermolecular condensation of the reactive groups, addition polymers ordinarily result from chain reactions involving some sort of active centres.
8. Chain polymerisation is characteristic of unsaturated monomers. Ethenic polymerisation is an economically important class whose kinetics typify chain polymerisation. The terms 'vinyl', 'olefin' or 'addition' polymerisation are often used, although they are more restrictive. Flory showed conclusively that chain polymerisation proceeds by and requires the steps of interaction, propagation, and termination typical of chain reactions in low-molecular-weight species.

1.8 SUMMARY

- Biophysics is the field that applies the theories and methods of physics to understand how biological systems work.
- Biophysics has been critical to understanding the mechanics of how the molecules of life are made, how different parts of a cell move and function, and how complex systems in our bodies—the brain, circulation, immune system, and others— work.
- Biophysics is, therefore, an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena.
- Biophysics covers all scales of biological organisation, from molecular to organismic and populations. Biophysical research shares significant overlap with biochemistry, molecular biology, physical chemistry, physiology, nanotechnology, bioengineering, computational biology, biomechanics, developmental biology and systems biology.
- Molecular biophysics typically statements biological questions similar to those in biochemistry and molecular biology, seeking to find the physical underpinnings of biomolecular phenomena. Scientists in this field conduct

NOTES

research concerned with understanding the interactions between the various systems of a cell, including the interactions between DeoxyriboNucleic Acid (DNA), RiboNucleic Acid (RNA) and protein biosynthesis, as well as how these interactions are regulated.

- Fluorescent imaging techniques, as well as Electron Microscopy, X-ray Crystallography, Nuclear Magnetic Resonance (NMR) spectroscopy, Atomic Force Microscopy (AFM) and Small-Angle Scattering (SAS) both with X-rays and Neutrons [Small-Angle Neutron Scattering (SANS)/ Small-Angle X-ray Scattering (SAXS)] are often used to visualise structures of biological significance. Protein dynamics can be observed by neutron spin echo spectroscopy.
- The field of quantum biology applies quantum mechanics to biological objects and problems. Decohered isomers to yield time-dependent base substitutions. These studies imply applications in quantum computing.
- Typically, an atom is the smallest unit of matter that retains all of the chemical properties of an element. Atoms combine to form molecules, which then interact to form solids, gases, or liquids. For example, water is composed of hydrogen and oxygen atoms that have combined to form water molecules.
- Several biological processes are dedicated to breaking down molecules into their component atoms so that they can be reconstructed into a more useful molecule.
- Atoms are made up of particles called protons, neutrons, and electrons, which are responsible for the mass and charge of atoms.
- Atoms consist of three basic particles: protons, electrons, and neutrons. The nucleus found at the centre in the place an atom whereas the protons contain positively charged and the neutrons have a no charge like a neutral.
- The outermost regions of the atom are called electron shells and contain the electrons negatively charged. Atoms have different properties based on the arrangement and number of their basic particles.
- The Hydrogen (H) atom contains only one proton, one electron, and no neutrons. This can be determined using the atomic number and the mass number of the element.
- Protons and neutrons have approximately the same mass, about 1.67×10^{-24} grams. Scientists define this amount of mass as one atomic mass unit (amu) or one Dalton. Although similar in mass, protons are positively charged, while neutrons have no charge. Therefore, the number of neutrons in an atom contributes significantly to its mass, but not to its charge.
- Chemical bonds are forces that hold the atoms together in a molecule. They are a result of strong intramolecular interactions among the atoms of a molecule.

- The valance (outermost) electrons of the atoms participate in chemical bonds. When two atoms approach each other, these outer electrons start to interact. Although electrons repel each other, they are attracted to the protons within atoms.
- The forces between the atoms are characterised by isotropic continuum electrostatic potentials. Their magnitude is in simple proportion to the charge difference.
- Covalent bonds are better understood by Valence Bond (VB) theory or Molecular Orbital (MO) theory. The properties of the atoms involved can be understood using concepts, such as oxidation number, formal charge, and electronegativity.
- Valence bond theory is more chemically intuitive by being spatially localised, allowing attention to be focused on the parts of the molecule undergoing chemical change. In contrast, molecular orbitals are more natural from a quantum mechanical point of view, with orbital energies being physically significant and directly linked to experimental ionisation energies from photoelectron spectroscopy.
- In the general case, atoms form bonds that are intermediate between ionic and covalent, depending on the relative electronegativity of the atoms involved. Bonds of this type are known as polar covalent bonds.
- The strength of chemical bonds varies considerably; there are 'Strong Bonds' or Primary Bonds', such as covalent, ionic and metallic bonds, and 'Weak Bonds' or 'Secondary Bonds', such as dipole–dipole interactions, the London dispersion force and hydrogen bonding.
- Covalent bonding is a common type of bonding in which two or more atoms share valence electrons more or less equally. The simplest and most common type is a single bond in which two atoms share two electrons. Other types include the double bond, the triple bond, one- and three-electron bonds, the three-centre two-electron bond and three-centre four-electron bond.
- A single bond between two atoms corresponds to the sharing of one pair of electrons. The Hydrogen (H) atom has one valence electron. Two Hydrogen atoms can then form a molecule, held together by the shared pair of electrons. Each H atom now has the noble gas electron configuration of Helium (He). The pair of shared electrons forms a single covalent bond. The electron density of these two bonding electrons in the region between the two atoms increases from the density of two non-interacting H atoms.
- Ionic bonding is a type of electrostatic interaction between atoms that have a large electronegativity difference. There is no precise value that distinguishes ionic from covalent bonding, but an electronegativity difference of over 1.7 is likely to be ionic while a difference of less than 1.7 is likely to be covalent.

NOTES

- Ionic bonding leads to separate positive and negative ions. Ionic charges are commonly between “3e to +3e. Ionic bonding commonly occurs in metal salts, such as Sodium Chloride (salt).
- Polymers are one of the most important products, which find an important place in every walk of modern civilisation. The term polymer (Greak word: poly + meros, means, many parts) denotes a molecule produced by the repetition of some simpler unit, called the mer or the monomer.
- Inorganic include polymers containing no carbon atoms. There are composed of different atoms joined by chemical bonds, while weaker intermolecular forces act between their chains. Polysilanes, polygermanes, polysilicic acid, polyphosphates, polyarsenates, etc., are examples of inorganic polymers.
- The reactions by which complex macromolecules are formed are not completely understood. The process of building up polymers from simple repeating units (monomers) can proceed with many variations. On the basis of mechanisms the two types of polymers are formed.
- Condensation polymers are usually formed by the step-wise intermolecular condensation of the reactive groups, addition polymers ordinarily result from chain reactions involving some sort of active centres.
- Chain polymerisation is characteristic of unsaturated monomers. Ethenic polymerisation is an economically important class whose kinetics typify chain polymerisation. The terms ‘vinyl’, ‘olefin’ or ‘addition’ polymerisation are often used, although they are more restrictive. Flory showed conclusively that chain polymerisation proceeds by and requires the steps of interaction, propagation, and termination typical of chain reactions in low-molecular-weight species.

1.9 KEY WORDS

- **Biophysics:** Biophysics is an interdisciplinary science that applies approaches and methods traditionally used in physics to study biological phenomena. Biophysics covers all scales of biological organisation, from molecular to organismic and populations.
- **Atoms:** Atoms are made up of particles called protons, neutrons, and electrons, which are responsible for the mass and charge of atoms. An atom is the smallest unit of matter that holds as all of the chemical properties of an element. Atoms combine to form molecules, which then interact to form solids, gases, or liquids.
- **Covalent bond:** A covalent bond is a chemical bond that involves the sharing of electron pairs between atoms. These electron pairs are termed shared

NOTES

pairs or bonding pairs, and the stable balance of attractive and repulsive forces between atoms, when they share electrons, is termed covalent bonding.

- **Ionic bonding:** Ionic bonding is a type of chemical bond that involves the electrostatic attraction between oppositely charged ions, and is the primary interaction occurring in ionic compounds. The ions are atoms that have lost one or more electrons (cations) and atoms that have gained one or more electrons (anions).
- **Hydrogen bonding:** A hydrogen bond (H-bond) is a primarily electrostatic force of attraction between a Hydrogen (H) atom which is covalently bound to a more electronegative atom or group, and another electronegative atom bearing a lone pair of electrons the hydrogen bond Acceptor (Ac).

1.10 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the concept of biophysics.
2. Give the definitions of atom and molecules.
3. Elaborate on the structure of the atoms.
4. What are the chemical bonds?
5. What do you mean by the ionic bonding?
6. Define the London dispersion force.
7. State about the hydrogen bonding.
8. What is polymer?
9. Elaborate on the polymerisation process.
10. Describe the radical polymerisation.

Long-Answer Questions

1. Discuss briefly the concept of biophysics with the help of examples.
2. Analyse the structure and properties of atoms and molecules giving appropriate examples.
3. Describe in detail about the chemical bonds.
4. Differentiate between the primary bonds and secondary bonds with the help of examples.
5. What is polymerisation process? Explain the addition and condensation polymerisation.

6. Explain in detail about the chemistry of polymerisation.
7. Briefly explain about the effects of various factors on polymerisation rate.

NOTES

1.11 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 2 THERMODYNAMICS

NOTES

Structure

2.0 INTRODUCTION

All physical and chemical transformations involve energy changes as kinetic energy, potential, chemical, mechanical, electrical and magnetic energy, etc. Thermodynamics deal with energy and its application to all phenomena in nature. It is applicable for macroscopic system not for molecular level or for microscopic system. Any process that occurs on its own without outside intervention is known as spontaneous process or any process which proceeds in forward direction without any external factor is known as spontaneous process. When using entropy change of a process for spontaneity it is necessary to understand definition of system, surrounding and the second law of thermodynamics.

Bioenergetics is a field in biochemistry and cell biology that concerns energy flow through living systems. This is an active area of biological research that includes the study of the transformation of energy in living organisms and the study of thousands of different cellular processes, such as cellular respiration and the many other metabolic and enzymatic processes that lead to production and utilisation of energy in forms, such as Adenosine Triphosphate (ATP) molecules. That is, the goal of bioenergetics is to describe how living organisms acquire and transform energy in order to perform biological work. The study of metabolic pathways is thus essential to bioenergetics.

NOTES

A coupling reaction in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst. In one important reaction type, a main group organometallic compound of the type R-M (R = Organic Fragment, M = Main Group Centre) reacts with an organic halide of the type R'-X with formation of a new carbon-carbon bond in the product R-R'. The most common type of coupling reaction is the cross coupling reaction.

A spontaneous process is in which it releases free energy and it moves to a lower, more thermodynamically Stable Energy State. Depending on the nature of the process, the free energy is determined differently as Gibb's free energy is applicable when process consider under constant pressure and temperature condition whereas Helmholtz free energy is used when process is considered under constant volume and temperature. A spontaneous process is characterized by decrease in system's free energy so that no need to be driven any outside source of energy.

In this unit, you will study about the laws of thermodynamics, principles and applications of thermodynamics, bioenergetics, coupling of chemical reactions, redox potential, NADP/NADPH, free energy.

2.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept of thermodynamics
- Explain the principle and applications of thermodynamics
- Discuss about the bioenergetics
- Interpret the coupling chemical reactions
- Define the redox potential
- Analyse the NADP/NADPH
- Elaborate on the free energy

2.2 LAWS OF THERMODYNAMICS

The first law of thermodynamics has certain limitations as illustrated below:

The first law establishes definite relationship between the heat absorbed and the work performed by a system in a given process. But it pits no restriction on the direction of the flow of heat. According to the first law, for example, it is not impossible to extract heat from ice by cooling it to a low temperature and then use it for wanning water. But it is known from experience that such a transfer of heat from a lower to a higher temperature is not possible without expenditure of energy

NOTES

that is without doing some external work. It's known, on the other hand, that heat flows spontaneously, that is, of its own accord, from a higher to a lower temperature.

According to the first law, the total energy of an isolated system remains constant during a specified change of state. But it does not tell whether a specified change of or a process including a chemical reaction can occur spontaneously, whether it is feasible?

The first law states that energy of one form can be converted into an equivalent amount of energy of another form. But it does not tell that heat energy cannot be completely converted into an equivalent amount of work. There is thus need for another law, i.e., the second law of thermodynamics.

The second law of thermodynamics helps us to determine the direction in which energy can be transferred. It also helps us to predict whether a given process or a chemical reaction can occur spontaneously, that is, its own accord. It also helps us to know the equilibrium conditions. The law is therefore, of great importance in chemistry.

It is known from experience that although various forms of energy can be completely transformed into one another, yet heat is a typical form of energy which cannot be transformed into work. The second law helps us to calculate the maximum fraction of heat that can be converted into work in a given process.

The second law of thermodynamics is needed because the first law of thermodynamics does not define the energy conversion process completely. The first law is used to relate and to evaluate the various energies involved in a process. However, no information about the direction of the process can be obtained by the application of the first law. Early in the development of the science of thermodynamics, investigators noted that while work could be converted completely into heat, the converse was never true for a cyclic process. Certain natural processes were also observed always to proceed in a certain direction (for example, heat transfer occurs from a hot to a cold body). The second law was developed as an explanation of these natural phenomena.

The second law of thermodynamics is used to determine the maximum efficiency of any process. A comparison can then be made between the maximum possible efficiency and the actual efficiency obtained.

With the second law of thermodynamics, the limitations imposed on any process can be studied to determine the maximum possible efficiencies of such a process and then a comparison can be made between the maximum possible efficiency and the actual efficiency achieved. One of the areas of application of the second law is the study of energy-conversion systems. For example, it is not possible to convert all the energy obtained from a nuclear reactor into electrical energy. There must be losses in the conversion process. The second law can be used to derive an expression for the maximum possible energy conversion efficiency taking those losses into account. Therefore, the second law denies the possibility

NOTES

of completely converting into work all of the heat supplied to a system operating in a cycle, no matter how perfectly designed the system may be.

The second law of thermodynamics can be stated in a number of ways. The original statements were concerned with the French engineer Sadi Carnot's analysis of performance of steam engines. There are two important statements of the second law; one is due to Lord Kelvin and other is due to Clausius. The statement due to Kelvin may be expressed as:

It is impossible for a cyclic process to take heat from a cold reservoir and convert it into work without at the same time transferring heat from a hot to a cold reservoir.

The statement due to Clausius is:

It is impossible to construct a machine, which is able to convey heat by a cyclic process from one reservoir at a lower temperature to another at higher temperature unless work is done on the machine by some outside agency.

A cycle is a process in which a system returns to its original state after a succession of steps. A more useful statement of the second law can be given in terms of a quantity which is also a function of state called, *entropy*.

2.3 PRINCIPAL AND APPLICATIONS OF THERMODYNAMICS

The laws of thermodynamics define a group of physical quantities, such as temperature, energy, and entropy that characterise thermodynamic systems in thermodynamic equilibrium. The laws also use various parameters for thermodynamic processes, such as thermodynamic work and heat, and establish relationships between them. They state empirical facts that form a basis of precluding the possibility of certain phenomena, such as perpetual motion. In addition to their use in thermodynamics, they are important fundamental laws of physics in general, and are applicable in other natural sciences.

Traditionally, thermodynamics has recognised three fundamental laws, simply named by an ordinal identification, the first law, the second law, and the third law. A more fundamental statement was later labelled as the zeroth law, after the first three laws had been established.

The ***zeroth law of thermodynamics*** defines thermal equilibrium and forms a basis for the definition of temperature: If two systems are each in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.

The ***first law of thermodynamics*** states that, when energy passes into or out of a system (as work, heat, or matter), the system's internal energy changes in according with the law of conservation of energy.

NOTES

The ***second law of thermodynamics*** states that in a natural thermodynamic process, the sum of the entropies of the interacting thermodynamic systems never decreases. Another form of the statement is that heat does not spontaneously pass from a colder body to a warmer body.

The ***third law of thermodynamics states*** that a system's entropy approaches a constant value as the temperature approaches absolute zero. With the exception of non-crystalline solids (glasses) the entropy of a system at absolute zero is typically close to zero.

The first and second law prohibit two kinds of perpetual motion machines, respectively: the perpetual motion machine of the first kind which produces work with no energy input, and the perpetual motion machine of the second kind which spontaneously converts thermal energy into mechanical work.

2.3.1 Zeroth Law of Thermodynamics

The zeroth law of thermodynamics provides for the foundation of temperature as an empirical parameter in thermodynamic systems and establishes the transitive relation between the temperatures of multiple bodies in thermal equilibrium. The law may be stated in the following form:

- If two systems are both in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.
- Though this version of the law is one of the most commonly stated versions, it is only one of a diversity of statements that are labelled as the 'Zeroth Law'. Some statements go further, so as to supply the important physical fact that temperature is one-dimensional and that one can conceptually arrange bodies in a real number sequence from colder to hotter.
- These concepts of temperature and of thermal equilibrium are fundamental to thermodynamics and were clearly stated in the nineteenth century. The name 'Zeroth Law' was invented by Ralph H. Fowler in the 1930s, long after the first, second, and third laws were widely recognised. The law allows the definition of temperature in a non-circular way without reference to entropy, its conjugate variable. Such a temperature definition is said to be 'Empirical'.

2.3.2 First Law of Thermodynamics

The first law of thermodynamics is a related to the law of conservation of energy, adapted for thermodynamic processes. In general, the conservation law states that the total energy of an isolated system is constant; energy can be transformed from one form to another, but can be neither created nor destroyed.

In a closed system (i.e. there is no transfer of matter into or out of the system), the first law states that the change in internal energy of the system

(ΔU_{system}) is equal to the difference between the heat supplied to the system *(Q)* and the work *(W)* done by the system on its surroundings.

$$\Delta U_{system} = Q - W$$

When two initially isolated systems are combined into a new system, then the total internal energy of the new system, U_{system}, will be equal to the sum of the internal energies of the two initial systems, U_1 and U_2:

$$U_{system} = U_1 + U_2$$

The First Law Encompasses Several Principles:

- The Conservation of energy, which says that energy can be neither created nor destroyed, but can only change one form to another form. A particular consequence of this is that the total energy of an isolated system does not change.
- The concept of internal energy and its relationship to temperature. If a system has a definite temperature, then its total energy has three distinguishable components, termed kinetic energy (energy due to the motion of the system as a whole), potential energy (energy resulting from an externally imposed force field), and internal energy. The establishment of the concept of internal energy distinguishes the first law of thermodynamics from the more general law of conservation of energy.

  $$E_{\text{total}} = \text{KE}_{\text{system}} + \text{PE}_{\text{system}} + U_{\text{system}}$$

- Work is a process of transferring energy to or from a system in ways that can be described by macroscopic mechanical forces acting between the system and its surroundings. The work done by the system can come from its overall kinetic energy, from its overall potential energy, or from its internal energy.

  For example, when a machine (not a part of the system) lifts a system upwards, some energy is transferred from the machine to the system. The system's energy increases as work is done on the system and in this particular case, the energy increase of the system is manifested as an increase in the system's gravitational potential energy. Work added to the system increases the potential energy of the system:

  When matter is transferred into a system that masses associated internal energy and potential energy are transferred with it.

  $$(u \, \Delta M)_{\text{in}} = \Delta U_{system}$$

  Where u denotes the internal energy per unit mass of the transferred matter, as measured while in the surroundings; and ΔM denotes the amount of transferred mass.

NOTES

- The flow of heat is a form of energy transfer. Heating is the natural process of moving energy to or from a system other than by work or the transfer of matter. In a diathermic system, the internal energy can only be changed by the transfer of energy as heat:

  $\Delta U_{system} = Q$

  Combining these principles leads to one traditional statement of the first law of thermodynamics: it is not possible to construct a machine which will perpetually output work without an equal amount of energy input to that machine. Or more briefly, a perpetual motion machine of the first kind is impossible.

2.3.3 Second Law of Thermodynamics

The second law of thermodynamics indicates the irreversibility of natural processes, and, in many cases, the tendency of natural processes to lead towards spatial homogeneity of matter and energy, and especially of temperature. It can be formulated in a variety of interesting and important ways. One of the simplest is the Clausius statement that heat does not spontaneously pass from a colder to a hotter body.

It implies the existence of a quantity called the entropy of a thermodynamic system. In terms of this quantity it implies that when two initially isolated systems in separate but nearby regions of space, each in thermodynamic equilibrium with itself but not necessarily with each other, are then allowed to interact, they will eventually reach a mutual thermodynamic equilibrium. The sum of the entropies of the initially isolated systems is less than or equal to the total entropy of the final combination. Equality occurs just when the two original systems have all their respective intensive variables (temperature, pressure) equal; then the final system also has the same values.

The second law is applicable to a wide variety of processes, both *reversible* and *irreversible*. According to the second law, in a reversible heat transfer, an element of heat transferred, δQ, is the product of the temperature *(T),* both of the system and of the sources or destination of the heat, with the increment *(dS)* of the system's conjugate variable, its entropy *(S):*

$$\delta Q = T\,dS$$

While reversible processes are a useful and convenient theoretical limiting case, all natural processes are irreversible. A prime example of this irreversibility is the transfer of heat by conduction or radiation. It was known long before the discovery of the notion of entropy that when two bodies, initially of different temperatures, come into direct thermal connection, then heat immediately and spontaneously flows from the hotter body to the colder one.

Entropy may also be viewed as a physical measure concerning the microscopic details of the motion and configuration of a system, when only the

macroscopic states are known. Such details are often referred to as disorder on a microscopic or molecular scale, and less often as dispersal of energy. For two given macroscopically specified states of a system, there is a mathematically defined quantity called the difference of information entropy between them. This defines how much additional microscopic physical information is needed to specify one of the macroscopically specified states, given the macroscopic specification of the other often a conveniently chosen reference state which may be presupposed to exist rather than explicitly stated. A final condition of a natural process always contains microscopically specifiable effects which are not fully and exactly predictable from the macroscopic specification of the initial condition of the process. This is why entropy increases in natural processes the increase tells how much extra microscopic information is needed to distinguish the initial macroscopically specified state from the final macroscopically specified state. Equivalently, in a thermodynamic process, energy spreads.

2.3.4 Third Law of Thermodynamics

The third law of thermodynamics can be stated as: A system's entropy approaches a constant value as its temperature approaches absolute zero.

At zero temperature, the system must be in the state with the minimum thermal energy, the ground state. The constant value (not necessarily zero) of entropy at this point is called the *residual entropy* of the system. Note that, with the exception of non-crystalline solids (e.g. glasses) the residual entropy of a system is typically close to zero. However, it reaches zero only when the system has a unique ground state (i.e. the state with the minimum thermal energy has only one configuration, or microstate). Microstates are used here to describe the probability of a system being in a specific state, as each microstate is assumed to have the same probability of occurring, so macroscopic states with fewer microstates are less probable. In general, entropy is related to the number of possible microstates according to the Boltzmann principle:

$$S = k_B \ ln \ \Omega$$

Where S is the entropy of the system, k_B Boltzmann's constant, and Ω the number of microstates. At absolute zero there is only 1 microstate possible ($\Omega=1$ as all the atoms are identical for a pure substance and as a result all orders are identical as there is only one combination) and $ln(1)=0$

2.3.5 Applications of Thermodynamics

- Sweating in a Crowded Room: In a crowded room, everybody (every person) starts sweating. The body starts cooling down by transferring the body heat to the sweat. Sweat evaporates adding heat to the room. Again, this happens due to the first and second law of thermodynamics in action. One thing to keep in mind, heat is not lost but transferred attaining equilibrium with maximum entropy.

NOTES

- The first law of thermodynamics proclaims constancy of the total energy of isolated system for all changes, taking place in this system: energy cannot be created or destroyed. According to the second law of thermodynamics in isolated system entropy is always increasing or remaining constant. All processes in the Universe are oriented to the equilibrium state. Nevertheless, biological systems, and, consequently, ecological systems create order from disorder, they create and support chemical and physical non-equilibrium state the basis they live on.
- According to the second law, heat always flows from a body at a higher temperature to a body at the lower temperature. This law is applicable to all types of heat engine cycles including Otto, Diesel, etc., for all types of working fluids used in the engines. This law has led to the progress of present-day vehicles.
- Another application of second law is refrigerators and heat pumps based on the Reversed Carnot Cycle. If you want to move heat from a body at a lower temperature to a body at a higher temperature, then you have to supply external work. In the original Carnot cycle, heat produces work while in the Reversed Carnot cycle work is provided to transfer heat from lower temperature reservoir to a higher temperature reservoir.
- Removing heat from the food items in the refrigerator and throwing it away to the higher temperature atmosphere does not happen automatically. We need to supply external work via the compressor to make this happen in the refrigerator.
- Air conditioner and heat pump follow the similar law of thermodynamics. The air conditioner removes heat from the room and maintains it at a lower temperature by throwing the absorbed heat into the atmosphere. The heat pump absorbs heat from the atmosphere and supplies it to the room which is cooler in winters.
- Melting of Ice Cube: Ice cubes in a drink absorb heat from the drink making the drink cooler. If we forget to drink it, after some time, it again attains room temperature by absorbing the atmospheric heat. All this happens as per the first and second law of thermodynamics.

2.4 BIOENERGETICS

Bioenergetics is a field in biochemistry and cell biology that concerns energy flow through living systems. This is an active area of biological research that includes the study of the transformation of energy in living organisms and the study of thousands of different cellular processes, such as cellular respiration and the many other metabolic and enzymatic processes that lead to production and utilisation of energy in forms, such as Adenosine Triphosphate (ATP) molecules. That is, the

goal of bioenergetics is to describe how living organisms acquire and transform energy in order to perform biological work. The study of metabolic pathways is thus essential to bioenergetics.

NOTES

Bioenergetics is the part of biochemistry concerned with the energy involved in making and breaking of chemical bonds in the molecules found in biological organisms. It can also be defined as the study of energy relationships and energy transformations and transductions in living organisms. The ability to harness energy from a variety of metabolic pathways is a property of all living organisms that contains earth science. Growth, development, anabolism and catabolism are some of the central processes in the study of biological organisms, because the role of energy is fundamental to such biological processes. Life is dependent on energy transformations; living organisms survive because of exchange of energy between living tissues/ cells and the outside environment. Some organisms, such as autotrophs, can acquire energy from sunlight (through Photosynthesis) without needing to consume nutrients and break them down. Other organisms, like heterotrophs, must intake nutrients from food to be able to sustain energy by breaking down chemical bonds in nutrients during metabolic processes, such as Glycolysis and the Citric Acid cycle. Importantly, as a direct consequence of the first law of thermodynamics, autotrophs and heterotrophs participate in a universal metabolic network by eating autotrophs (plants), heterotrophs harness energy that was initially transformed by the plants during photosynthesis.

In a living organism, chemical bonds are broken and made as part of the exchange and transformation of energy. Energy is available for work (such as mechanical work) or for other processes (such as chemical synthesis and anabolic processes in growth), when weak bonds are broken and stronger bonds are made. The production of stronger bonds allows release of usable energy.

Adenosine Triphosphate (ATP) is the main 'Energy Currency' for organisms; the goal of metabolic and catabolic processes are to synthesize ATP from available starting materials (from the environment), and to break- down ATP (into Adenosine DiPhosphate (ADP) and inorganic phosphate) by utilising it in biological processes. In a cell, the ratio of ATP to ADP concentrations is known as the 'Energy Charge' of the cell. A cell can use this energy charge to relay information about cellular needs; if there is more ATP than ADP available, the cell can use ATP to do work, but if there is more ADP than ATP available, the cell must synthesize ATP via oxidative phosphorylation.

Living organisms produce ATP from energy sources, mostly sunlight or O_2, mainly via oxidative phosphorylation. The terminal phosphate bonds of ATP are relatively weak compared with the stronger bonds formed when ATP is hydrolysed (broken down by water) to adenosine diphosphate and inorganic phosphate. Here, it is the thermodynamically favourable free energy of hydrolysis that results in energy release; the phosphoanhydride bond between the terminal phosphate group and the rest of the ATP molecule does not itself contain this energy. An organism's

stockpile of ATP is used as a battery to store energy in cells. Utilisation of chemical energy from such molecular bond rearrangement powers biological processes in every biological organism.

NOTES

Living organisms obtain energy from organic and inorganic materials; i.e., ATP can be synthesised from a variety of biochemical precursors. For example, lithotrophs can oxidize minerals, such as nitrites or forms of sulphur, such as elemental sulphur, sulphites, and hydrogen sulphide to produce ATP. In photosynthesis, autotrophs produce ATP using light energy, whereas heterotrophs must consume organic compounds, mostly including carbohydrates, fats, and proteins. The amount of energy actually obtained by the organism is lower than the amount released in combustion of the food; there are losses in digestion, metabolism, and thermogenesis.

Environmental materials that an organism intakes are generally combined with oxygen to release energy, although some can also be oxidised anaerobically by various organisms. The bonds holding the molecules of nutrients together and in particular the bonds holding molecules of free oxygen together are relatively weak compared with the chemical bonds holding carbon dioxide and water together. The utilisation of these materials is a form of slow combustion because the nutrients are reacted with oxygen (the materials are oxidised slowly enough that the organisms do not actually produce fire). The oxidation releases energy because stronger bonds (bonds within water and carbon dioxide) have been formed. This net energy may evolve as heat, which may be used by the organism for other purposes, such as breaking other bonds to do chemistry required for survival.

Types of Reactions

An ***exergonic reaction*** is a spontaneous chemical reaction that releases energy. It is thermodynamically favoured, indexed by a negative value of ΔG (Gibbs free energy). Over the course of a reaction, energy needs to be put in, and this activation energy drives the reactants from a stable state to a highly energetically unstable transition state to a more stable state that is lower in energy. The reactants are usually complex molecules that are broken into simpler products. The entire reaction is usually catabolic. The release of energy (specifically of Gibbs free energy) is negative (i.e. $\Delta G < 0$) because the energy of the reactants is higher than that of the products.

An ***endergonic reaction*** is an anabolic chemical reaction that consumes energy. It is the opposite of an exergonic reaction. It has a positive ΔG, for instance because $\Delta H > 0$, which means that it takes more energy to break the bonds of the reactant than the energy of the products offer, i.e., the products have weaker bonds than the reactants. Thus, endergonic reactions are thermodynamically unfavourable and will not occur on their own at constant temperature. Additionally, endergonic reactions are usually anabolic.

The free energy gained or lost *(ΔG)* in a reaction can be calculated as follows:

$\Delta G = \Delta H - T\Delta S$

Where *"G* = Gibbs Free Energy Change

ΔH = Enthalpy Change

T = Temperature (in Kelvins)

ΔS = Entropy Change

NOTES

2.5 COUPLING OF CHEMICAL REACTION

A ***coupling reaction*** in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst. In one important reaction type, a main group organometallic compound of the type R-M (R = organic fragment, M = main group centre) reacts with an organic halide of the type R'-X with formation of a new carbon-carbon bond in the product R-R'. The most common type of coupling reaction is the cross coupling reaction.

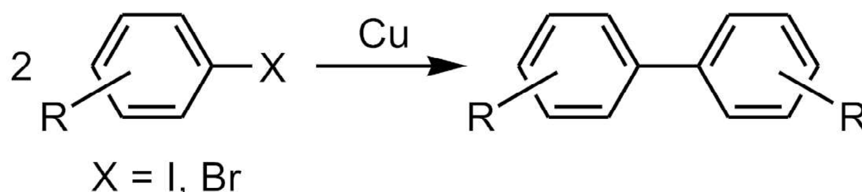
Richard F. Heck, Ei-ichi Negishi, and Akira Suzuki were awarded the 2010 Nobel Prize in Chemistry for developing palladium-catalysed cross coupling reactions.

Two types of coupling reactions are recognised:

- **Heterocouplings:** Heterocouplings combine two different partners, such as in the ***Heck reaction*** of an alkene (RC=CH) and an alkyl halide (R'-X) to give a substituted alkene. Heterocouplings are called ***cross-couplings***.
- **Homocouplings:** Homocouplings couple two identical partners, as in the ***Glaser coupling*** of two acetylides (RCa"CH) to form a dialkyne (RCa"C-Ca"CR).

Homo-Coupling Types

Coupling reactions are illustrated by the famous Ullmann reaction:



Reaction	Year	Reactant A		Reactant B		Reagent	Remark
Wurtz Reaction	1855	R-X	sp^3	R-X	sp^3	Na As Reducing Agent	
Pinacol Coupling Reaction	1859	R-HC=O or R_2(C=O)		R-HC=O or R_2(C=O)		Various Metals	Requires Proton Donor
Glaser Coupling	1869	RC≡CH	sp	RC≡CH	sp	Cu	O_2 as H-Acceptor
Ullmann Reaction	1901	Ar-X	sp^2	Ar-X	sp^2	Cu	High Temperatures

Cross-Coupling Types

NOTES

An illustrative cross-coupling reaction is the Heck coupling of an alkene and an aryl halide:

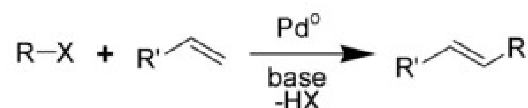


Table 2.1 *Different Types of Coupling Reactions*

Reaction	Year	Reactant A		Reactant B		Catalyst	Remark
Grignard Reaction	1900	R-MgBr	sp, sp^2, sp^3	R-HC=O or $R(C=O)R_2$	sp^2	Not Catalytic	
Gomberg-Bachmann Reaction	1924	Ar-H	sp^2	$Ar'\text{-}N_2^+X^-$	sp^2	Not Catalytic	
Cadiot-Chodkiewicz Coupling	1957	RC≡CH	sp	RC≡CX	sp	Cu	Requires Base
Castro-Stephens Coupling	1963	RC≡CH	sp	Ar-X	sp^2	Cu	
Corey-House Synthesis	1967	R_2CuLi or RMgX	sp^3	R-X	sp^2, sp^3	Cu	Cu-Catalyzed Version by Kochi, 1971
Cassar Reaction	1970	Alkene	sp^2	R-X	sp^3	Pd	Requires Base
Kumada Coupling	1972	Ar-MgBr	sp^2, sp^3	Ar-X	sp^2	Pd or Ni or Fe	
Heck Reaction	1972	alkene	sp^2	Ar-X	sp^2	Pd or Ni	Requires Base
Sonogashira Coupling	1975	RC≡CH	sp	R-X	sp^3 sp^2	Pd and Cu	Requires Base
Negishi Coupling	1977	R-Zn-X	sp^3, sp^2, sp	R-X	sp^3 sp^2	Pd or Ni	
Stille Cross Coupling	1978	$R\text{-}SnR_3$	sp^3, sp^2, sp	R-X	sp^3 sp^2	Pd	
Suzuki Reaction	1979	$R\text{-}B(OR)_2$	sp^2	R-X	sp^3 sp^2	Pd or Ni	Requires Base
Hiyama Coupling	1988	$R\text{-}SiR_3$	sp^2	R-X	sp^3 sp^2	Pd	Requires Base
Buchwald-Hartwig Reaction	1994	$R_2N\text{-}H$	sp^3	R-X	sp^2	Pd	N-C Coupling, Second Generation free Amine
Fukuyama Coupling	1998	R-Zn-I	sp^3	RCO(SEt)	sp^2	Pd or Ni[6]	
Liebeskind–Srogl Coupling	2000	$R\text{-}B(OR)_2$	sp^3, sp^2	RCO(SEt) Ar-SMe	sp^2	Pd	Requires Cu tc

NOTES

2.6 REDOX POTENTIAL

Redox potential also known as ***Oxidation Reduction Potential (ORP),*** the calculate value of redox potentials equal to the pe, E_0', or E_h is a measure of the tendency of a chemical species to acquire electrons from or lose electrons to an electrode and in this manner be ***reduced or oxidised*** respectively. Redox potential is measured in Volts (V), or milli Volts (mV). Each species has its own intrinsic redox potential, for example the more positive the reduction potential (reduction potential is more often used due to general formalism in electrochemistry), the greater the species' affinity for electrons and tendency to be reduced. ORP can reflect the antimicrobial potential of the water.

Measurement and Interpretation of Redox Potential

In aqueous solutions, redox potential is a measure of tendency of the solution to either gain or lose electrons when it is subjected to change by introduction of a new species. A solution with a higher (more positive) reduction potential than the new species will have a tendency to gain electrons from the new species (i.e. to be reduced by oxidising the new species) and a solution with a lower (more negative) reduction potential will have a tendency to lose electrons to the new species (i.e. to be oxidised by reducing the new species). Because the absolute potentials are next to impossible to accurately measure, reduction potentials are defined relative to a ***reference electrode***. Reduction potentials of aqueous solutions are determined by measuring the potential difference between an inert sensing electrode in contact with the solution and a stable reference electrode connected to the solution by a salt bridge.

The sensing electrode acts as a platform for electron transfer to or from the reference half-cell; it is typically made of platinum, although gold and graphite can be used as well. The reference half-cell consists of a redox standard of known potential. The Standard Hydrogen Electrode (SHE) is the reference from which all standard redox potentials are determined, and has been assigned an arbitrary half-cell potential of 0.0 mV. However, it is fragile and impractical for routine laboratory use. Therefore, other more stable reference electrodes, such as silver chloride and Saturated Calomel Electrodes (SCE) are commonly used because of the act is more reliable.

Although measurement of the redox potential in aqueous solutions is relatively straightforward, many factors limit its interpretation, such as effects of solution temperature and pH, ***irreversible reactions***, slow electrode kinetics, non-equilibrium, presence of multiple redox couples, electrode poisoning, small exchange currents, and inert redox couples. Consequently, practical measurements seldom correlate with calculated values. Nevertheless, reduction potential measurement has proven useful as an analytical tool in monitoring changes in a system rather than determining their absolute value (e.g. process control and titrations).

NOTES

Similar to how the concentration of hydrogen ion determines the acidity or pH of an aqueous solution, the tendency of electron transfer between a chemical species and an electrode determines the redox potential of an electrode couple. Like pH, redox potential represents how easily electrons are transferred to or from species in solution. Redox potential characterises the ability under the specific condition of a chemical species to lose or gain electrons instead of the amount of electrons available for oxidation or reduction.

In fact, it is possible to define pe, the negative logarithm of electron concentration ("log [e^-]) in a solution, which will be directly proportional to the redox potential. Sometimes pe is used as a unit of reduction potential instead of E_h, for example, in environmental chemistry. If we normalise pe of hydrogen to zero, we will have the relation pe = 16.9 E_h, at room temperature. This point of view is useful for understanding redox potential, although the transfer of electrons, rather than the absolute concentration of free electrons in thermal equilibrium, is how one usually thinks of redox potential. Theoretically, however, the two approaches are equivalent.

On the other hand, one could define a potential corresponding to pH as a potential difference between a solute and pH neutral water, separated by porous membrane (that is permeable to hydrogen ions). Such potential differences actually do occur from differences in acidity on biological membranes. This potential (where pH neutral water is set to 0 V) is analogous with redox potential (where standardised hydrogen solution is set to 0 V), but instead of hydrogen ions, electrons are transferred across in the redox case. Both pH and redox potentials are properties of solutions, not of elements or chemical compounds themselves, and depend on concentrations, temperature, etc.

2.7 NADP/NADPH

Nicotinamide Adenine Dinucleotide (NAD) is a coenzyme central to metabolism. Found in all living cells, NAD is called a dinucleotide because it consists of two nucleotides joined through their phosphate groups. One nucleotide contains an adenine nucleobase and the other nicotinamide. NAD exists in two forms: an oxidised and reduced form, abbreviated as NAD^+ and NADH (H for hydrogen) respectively. Figure 2.1 illustrate the structure of Nicotinamide Adenine Dinucleotide (NAD).

NOTES

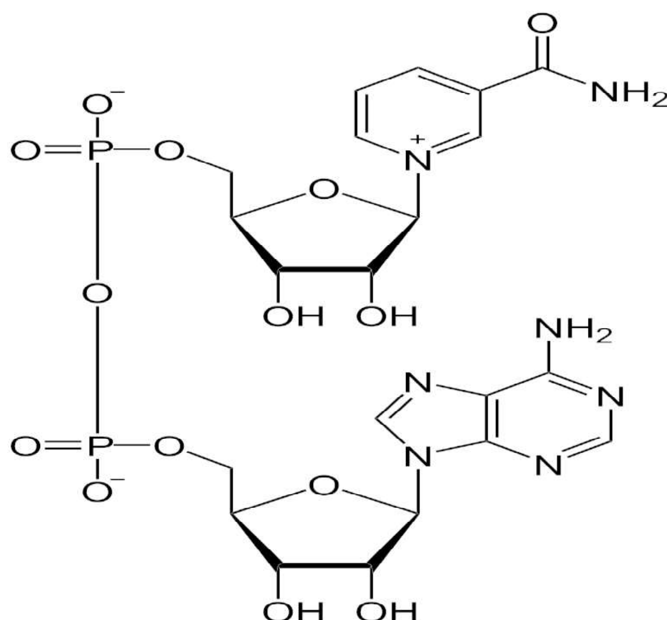


Fig. 2.1 *Structure of Nicotinamide Adenine Dinucleotide (NAD)*

In metabolism, nicotinamide adenine dinucleotide is involved in redox reactions, carrying electrons from one reaction to another. The cofactor is, therefore, found in two forms in cells: NAD^+ is an oxidising agent it accepts electrons from other molecules and becomes reduced. This reaction forms NADH, which can then be used as a reducing agent to donate electrons. These electron transfer reactions are the main function of NAD. However, it is also used in other cellular processes, most notably as a substrate of enzymes in adding or removing chemical groups to or from, respectively, proteins, in posttranslational modifications. Because of the importance of these functions, the enzymes involved in NAD metabolism are targets for drug discovery.

In organisms, NAD can be synthesised from simple building-blocks from either tryptophan or aspartic acid, each a case of an amino acid; alternatively, more complex components of the coenzymes are taken up from nutritive compounds, such as niacin; similar compounds are produced by reactions that break down the structure of NAD, providing a salvage pathway that 'Recycles' them back into their respective active form.

Some NAD is converted into the coenzyme Nicotinamide Adenine Dinucleotide Phosphate (NADP); its chemistry largely parallels that of NAD, though predominantly its role is as a cofactor in anabolic metabolism.

The NAD^+ chemical species superscripted addition sign reflects the formal charge on one of its nitrogen atoms; this species is actually a singly charged anion carrying a (negative) ionic charge of 1 under conditions of physiological pH. NADH, in contrast, is a doubly charged anion.

NOTES

Nicotinamide Adenine Dinucleotide Phosphate ($NADP^+$) Figure 2.2 illustrate the structure of ($NADP^+$) or in older notation, TPN (TriPhosphoPyridine Nucleotide), is a cofactor used in anabolic reactions, such as the Calvin cycle and lipid and nucleic acid syntheses, which require NADPH as a reducing agent. It is used by all forms of cellular life.

NADPH is the reduced form of $NADP^+$. $NADP^+$ differs from NAD^+ by the presence of an additional phosphate group on the 2' position of the ribose ring that carries the adenine moiety. This extra phosphate is added by NAD^+ kinase and removed by $NADP^+$ phosphatase.

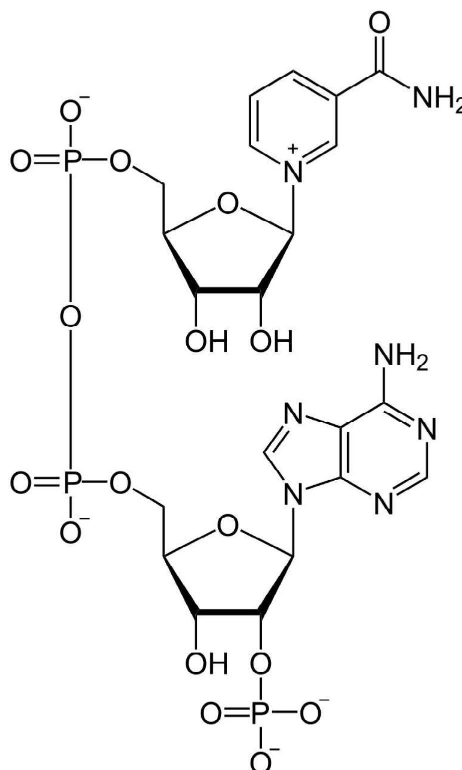


Fig.2.2 *Structure of Nicotinamide Adenine Dinucleotide Phosphate*

Biosynthesis

$NADP^+$: In general, $NADP^+$ is synthesized before NADPH is. Such a reaction usually starts with NAD^+ from either the de-novo or the salvage pathway, with NAD^+ kinase adding the extra phosphate group. $NAD(P)^+$ nucleosidase allows for synthesis from nicotinamide in the salvage pathway, and $NADP^+$ phosphatase can convert NADPH back to NADH to maintain a balance. Some forms of the NAD^+ kinase, notably the one in mitochondria, can also accept NADH to turn it directly into NADPH. The prokaryotic pathway is less well understood, but with all the similar proteins the process should work in a similar way.

NOTES

NADPH: NADPH is produced from $NADP^+$. The major source of NADPH in animals and other non-photosynthetic organisms is the pentose phosphate pathway, by Glucose-6-Phosphate Dehydrogenase (G6PDH) in the first step. The pentose phosphate pathway also produces pentose, another important part of NAD(P)H, from glucose.

Ferredoxin-$NADP^+$ reductase, present in all domains of life, is a major source of NADPH in photosynthetic organisms including plants and cyanobacteria. It appears in the last step of the electron chain of the light reactions of photosynthesis. It is used as reducing power for the biosynthetic reactions in the Calvin cycle to assimilate carbon dioxide and help turn the carbon dioxide into glucose. It has functions in accepting electrons in other non-photosynthetic pathways as well: it is needed in the reduction of nitrate into ammonia for plant assimilation in nitrogen cycle and in the production of oils.

There are several other lesser-known mechanisms of generating NADPH, all of which depend on the presence of mitochondria in eukaryotes. The key enzymes in these carbon-metabolism-related processes are NADP-linked isoforms of malic enzyme, Isocitrate Dehydrogenase (IDH), and glutamate dehydrogenase. In these reactions, $NADP^+$ acts like NAD^+ in other enzymes as an oxidising agent. The isocitrate dehydrogenase mechanism appears to be the major source of NADPH in fat and possibly also liver cells. These processes are also found in bacteria. Bacteria can also use a NADP-dependent glyceraldehyde 3-phosphate dehydrogenase for the same purpose. Like the pentose phosphate pathway, these pathways are related to parts of glycolysis.

NADPH can also be generated through pathways unrelated to carbon metabolism. The ferredoxin reductase is such an example. Nicotinamide Nucleotide transhydrogenase transfers the Hydrogen between NAD(P)H and $NAD(P)^+$, and is found in eukaryotic mitochondria and many bacteria. There are versions that depend on a proton gradient to work and ones that do not. Some anaerobic organisms use $NADP^+$-linked hydrogenase, ripping a hydride from hydrogen gas to produce a proton and NADPH.

Like NADH, NADPH is fluorescent. NADPH in aqueous solution excited at the nicotinamide absorbance of ~335 nm (near UV) has a fluorescence emission which peaks at 445-460 nm (violet to blue). $NADP^+$ has no appreciable fluorescence.

2.8 FREE ENERGY

Free energy, in thermodynamics, energy-like property or state function of a system in thermodynamic equilibrium. Free energy has the dimensions of energy, and its value is determined by the state of the system and not by its history. Free energy is

used to determine how systems change and how much work they can produce. It is expressed in following two forms:

- The Helmholtz free energy F, sometimes called the work function
- The Gibbs free energy G

If U is the internal energy of a system, PV the pressure-volume product, and TS the temperature-entropy product (T being the temperature above absolute zero), then F = U – TS and G = U + PV – TS. The latter equation can also be written in the form G = H – TS, where H = U + PV is the enthalpy. Free energy is an extensive property, meaning that its magnitude depends on the amount of a substance in a given thermodynamic state.

The changes in free energy, ΔF or ΔG, are useful in determining the direction of spontaneous change and evaluating the maximum work that can be obtained from thermodynamic processes involving chemical or other types of reactions. In a reversible process the maximum useful work that can be obtained from a system under constant temperature and constant volume is equal to the (negative) change in the Helmholtz free energy, $-\Delta F = -\Delta U + T\Delta S$, and the maximum useful work under constant temperature and constant pressure (other than work done against the atmosphere) is equal to the (negative) change in the Gibbs free energy, $-\Delta G = -\Delta H + T\Delta S$. In each case, the $T\Delta S$ entropy term represents the heat absorbed by the system from a heat reservoir at temperature T under conditions where the system does maximum work. By conservation of energy, the total work done also includes the decrease in internal energy U or enthalpy H as the case may be. For example, the energy for the maximum electrical work done by a battery as it discharges comes both from the decrease in its internal energy due to chemical reactions and from the heat $T\Delta S$ it absorbs in order to keep its temperature constant, which is the ideal maximum heat that can be absorbed. For any actual battery, the electrical work done would be less than the maximum work, and the heat absorbed would be correspondingly less than $T\Delta S$.

Changes in free energy can be used to judge whether changes of state can occur spontaneously. Under constant temperature and volume, the transformation will happen spontaneously, either slowly or rapidly, if the Helmholtz free energy is smaller in the final state than in the initial state, that is, if the difference ÄF between the final state and the initial state is negative. Under constant temperature and pressure, the transformation of state will occur spontaneously if the change in the Gibbs free energy, ΔG, is negative.

Phase transitions provide instructive examples, as when ice melts to form water at 0.01 °C (T = 273.16 K), with the solid and liquid phases in equilibrium. Then ΔH = 79.71 calories per gram is the latent heat of fusion, and by definition $\Delta S = \Delta H/T$ = 0.292 calories per gram'" K is the entropy change. It follows immediately that $\Delta G = \Delta H - T\Delta S$ is zero, indicating that the two phases are in equilibrium and that no useful work can be extracted from the phase transition (other than work against the atmosphere due to changes in pressure and volume).

Furthermore, ΔG is negative for T > 273.16 K, indicating that the direction of spontaneous change is from ice to water, and ΔG is positive for T < 273.16 K, where the reverse reaction of freezing takes place.

Gibb's Free Energy

Gibbs free energy is a measure of the potential for reversible or maximum work that may be done by a system at constant temperature and pressure. It is a thermodynamic property that was defined in 1876 by Josiah Willard Gibbs to predict whether a process will occur spontaneously at constant temperature and pressure. Gibbs free energy G is defined as G = H – TS where H, T, and S are the enthalpy, temperature, and entropy. The SI unit for Gibbs energy is the kilo joule (kJ).

Changes in the Gibbs free energy G correspond to changes in free energy for processes at constant temperature and pressure. The change in Gibbs free energy change is the maximum non-expansion work obtainable under these conditions in a closed system. ΔG is negative for spontaneous processes, positive for nonspontaneous processes and zero for processes at equilibrium.

Also Known As: (G), Gibbs' free energy, Gibbs energy, or Gibbs function. Sometimes the term 'free enthalpy' is used to distinguish it from Helmholtz free energy. The terminology recommended by the International Union of Pure and Applied Chemistry (IUPAC) is Gibbs energy or Gibbs function.

Positive and Negative Free Energy

The sign of a Gibbs energy value may be used to determine whether or not a chemical reaction proceeds spontaneously. If the sign for ΔG is positive, additional energy must be input for the reaction to occur. If the sign for ΔG is negative, the reaction is thermodynamically favorable and will occur spontaneously. However, just because a reaction occurs spontaneously doesn't mean it occurs quickly. The formation of rust (iron oxide) from iron is spontaneous, yet occurs too slowly to observe. The reaction C(s) diamond → C(s) graphite also has a negative ΔG at 25 °C and 1 atm, yet diamonds are not seen to spontaneously change into graphite.

The Gibbs free energy is defined as: *TS*

$$G(p,T) = U + pV - TS,$$

which is the same as

$$G(p,T) = H - TS,$$

Where:

U is the internal energy (SI unit: Joule)

p is pressure (SI unit: Pascal)

V is volume (SI unit: m^3)

T is the temperature (SI unit: Kelvin)

S is the entropy (SI unit: Joule per Kelvin)

H is the enthalpy (SI unit: Joule)

NOTES

The expression for the infinitesimal reversible change in the Gibbs free energy as a function of its 'natural variables' p and T, for an open system, subjected to the operation of external forces (for instance, electrical or magnetic) X_i, which cause the external parameters of the system a_i to change by an amount da_i, can be derived as follows from the first law for reversible processes:

$$T\,dS = dU + p\,dV - \sum_{i=1}^{k} \mu_i\,dN_i + \sum_{i=1}^{n} X_i\,da_i + \cdots$$

$$d(TS) - S\,dT = dU + d(pV) - V\,dp - \sum_{i=1}^{k} \mu_i\,dN_i + \sum_{i=1}^{n} X_i\,da_i + \cdots$$

$$d(U - TS + pV) = V\,dp - S\,dT + \sum_{i=1}^{k} \mu_i\,dN_i - \sum_{i=1}^{n} X_i\,da_i + \cdots$$

$$dG = V\,dp - S\,dT + \sum_{i=1}^{k} \mu_i\,dN_i - \sum_{i=1}^{n} X_i\,da_i + \cdots$$

Where:

u_i is the chemical potential of the $_i$th chemical component. (SI unit: Joules per particle or Joules per mole)

N_i is the number of particles (or number of moles) composing the $_i$th chemical component.

The temperature dependence of the Gibbs energy for an ideal gas is given by the Gibbs–Helmholtz equation, and its pressure dependence is given by,

$$\frac{G}{N} = \frac{G^\circ}{N} + kT \ln \frac{p}{p^\circ}.$$

If the volume is known rather than pressure, then it becomes

$$\frac{G}{N} = \frac{G^\circ}{N} + kT \ln \frac{V^\circ}{V},$$

or more conveniently as its chemical potential:

$$\frac{G}{N} = \mu = \mu^\circ + kT \ln \frac{p}{p^\circ}.$$

In non-ideal systems, fugacity comes into play.

Helmholtz Free Energy

In thermodynamics, the Helmholtz free energy is a thermodynamic potential that measures the useful work obtainable from a closed thermodynamic system at a constant temperature and volume (isothermal, isochoric). The negative of the change in the Helmholtz energy during a process is equal to the maximum amount

NOTES

of work that the system can perform in a thermodynamic process in which volume is held constant. If the volume were not held constant, part of this work would be performed as boundary work. This makes the Helmholtz energy useful for systems held at constant volume. Furthermore, at constant temperature, the Helmholtz energy is minimized at equilibrium.

In contrast, the Gibbs free energy or free enthalpy is most commonly used as a measure of thermodynamic potential (especially in chemistry) when it is inconvenient for applications that do not occur at constant pressure. For example, in explosives research Helmholtz free energy is often used, since explosive reactions by their nature induce pressure changes. It is also frequently used to define fundamental equations of state of pure substances.

The concept of free energy was developed by Hermann von Helmholtz, a German physician and physicist, and first presented in 1882 in a lecture called 'On the thermodynamics of chemical processes'. From the German word Arbeit (work), the International Union of Pure and Applied Chemistry (IUPAC) recommends the symbol A and the name Helmholtz energy. In physics, the symbol F is also used in reference to free energy or Helmholtz function.

The Helmholtz energy is defined as:

$$F \equiv U - TS,$$

Where

F is the Helmholtz free energy (SI: Joules, CGS: Ergs),

U is the internal energy of the system (SI: Joules, CGS: Ergs),

T is the absolute temperature (Kelvins) of the surroundings, modelled as a heat bath,

S is the entropy of the system (SI: Joules per Kelvin, CGS: Ergs per Kelvin).

The Helmholtz energy is the Legendre transformation of the internal energy *U*, in which temperature replaces entropy as the independent variable.

Criterion of Equilibrium

The second law of thermodynamics gives the condition of maximum entropy to define the position of equilibrium in systems at constant energy. This is a useful criterion for spontaneous chemical reactions.

Consider a system in contact with a reservoir at temperature T in which an infinitesimal irreversible process occurs and the only work done is the pressure-volume work. If q is the quantity of heat exchanged with the reservoir, then, since the process is irreversible, the entropy change dS for the system is greater than $\frac{q}{T}$, i.e.,

$$dS > \frac{q}{T} \qquad ...(1)$$

or $$TdS > q$$

Since TdS is greater than q, $(q - TdS)$ is negative, i.e., less than zero

$$q - TdS < 0 \qquad ...(2)$$

Since the only work is the pressure-volume work,

NOTES

$$q = dE + pdV$$

and Equation (2) becomes

$$dE + pdV - TdS < 0 \qquad ...(3)$$

This inequality is always applicable if a spontaneous change occurs and the only work involved is the pressure-volume work. If the volume and entropy of the system are held constant, then equation (3) reduces to

$$(dE)_{v,\ s} < 0 \qquad ...(4)$$

Thus, for any irreversible process in a system of constant volume that does not change its entropy, the internal energy decreases. In other words, for a conservative mechanical system, the stable state is the one of lowest energy. A more careful statement of the role of the internal energy on position of equilibrium is that in systems of constant entropy. The equilibrium position is in the direction of lowest energy.

The volume and the internal energy of a system may be taken constant by isolating the system. For such a system, Equation (3) becomes:

$$(-TdS)_{E,\ V} < 0$$

which when multiplied by –1 and divided by T gives

$$(dS)_{V,E} > 0 \qquad ...(5)$$

and so the entropy must increase in such an irreversible process. It can also be put in words that in systems of constant energy and constant volume, such as isolated systems, the equilibrium position is in the direction of highest entropy.

For the system which is not isolated, there are entropy changes in the adjacent systems which must also be considered. If the volume is constant during the infinitesimal irreversible process, Equation (3) becomes:

$$(dE - TdS)\ v < 0 \qquad ...(6)$$

which can also be written as

$$d(E - TS)_{T,}\ v < 0 \qquad ...(7)$$

The quantity $(E - TS)$ is referred to as the *Helmholtz free energy* and is represented by A,

$$A = E - TS \qquad ...(8)$$

Differentiating this equation at constant temperature

$$dA = dE - TdS \qquad ....(9)$$

and for a spontaneous process, Equation (9) then becomes

$$(dA)_{T,V} < 0 \qquad ...(10)$$

In other words, in an irreversible process at constant T and V, the Helmholtz free energy A decreases.

NOTES

For a system at constant T and V, the PV work term will be zero. The system does no other work, therefore,

$$dA = 0 \qquad ...(11)$$

Hence, the condition of equilibrium for a system that can do no work is

$$dA = 0.$$

To a chemist, the physical processes or chemical reactions which are of interest are usually carried out in the laboratory at constant temperature and pressure. For such processes, Equation (3) can be written as:

$$d(E + PV - TS)_{T,P} < 0 \qquad ...(3a)$$

The quantity $(E + PV - TS)$ is referred to as the *Gibb's free energy* and is represented by the symbol G, i.e.,

$$G = E + PV - TS \qquad ...(12)$$

or $$G = H - TS \qquad (\because H = E + PV)$$

For infinitesimal change under isothermal conditions, (i.e., $dT = 0$)

$$dG = dH - TdS \qquad (\because SdT = 0) \qquad ...(13)$$

This is an important relationship and can be put into words:

For a change at constant pressure and constant temperature and where only mechanical work is done, the free energy is equal to the enthalpy change minus the product of the absolute temperature and the entropy change.

Equation 13 can be written as

$$(dG)_{T,P} < 0 \qquad ...(14)$$

Thus, in an irreversible process at constant T and P in which only pressure volume work is done, the Gibb's free energy decreases. Since both G and A are defined by an explicit equation in terms of variables which depend only upon the state of a system, both these are thermodynamic properties and their differentials are complete differential. Mathematically for a reversible cyclic process, we may write,

$$\oint dG = 0 \qquad ...(15)$$

$$\oint dA = 0 \qquad ...(16)$$

Table 2.2 summarizes the conditions for reversibility and irreversibility for processes involving only pressure-volume work.

NOTES

Table 2.2 *Criteria for Irreversibility and Reversibility for Processes Involving only Pressure-Volume Work*

For irreversible process	*For reversible process*
$(dS)_{E,v} > 0$	$(dS)_{v,E} = 0$
$(dS)_{v,S} < 0$	$(dS)_{v,E} = 0$
$(dA)_{T,\ v} < 0$	$(dA)_{T,\ v} = 0$
$(dG)_{T,\ P} < 0$	$(dG)_{T,\ P} = 0$

These relations may be applied to finite changes as well as to infinitesimal changes by replacing the *ds* by Δs. It must however be remembered that spontaneous changes always go to the minimum (as in the case of Gibb's free energy at constant *T* and *P*) or to the maximum (as in the case of the entropy of an isolated system) and not to some other conditions even though the change to some other conditions satisfies the required inequality.

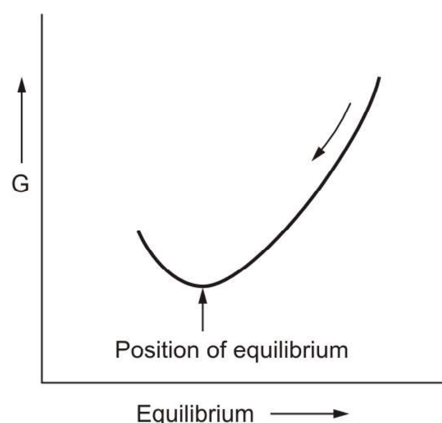


Fig. 2.3 *Position of Equilibrium in Terms of Gibb's Free Energy for a System at Constant Temperature and Pressure*

In this discussion, we have restricted ourselves only to those systems which involve pressure-volume work. If electrical work is done by the system, or if the temperature changes, the criteria for equilibrium will be altered. If a system is at constant *T* and *P* and does no additional work, the condition of equilibrium is $dG = 0$, *G* will be at minimum when the system is at equilibrium as can be seen from Figure 2.3.

Free Energy and Boiling Points

It is interesting to consider why a liquid boils at a particular temperature. Why, for example, does water at 1 atm pressure boil at 373 °K.

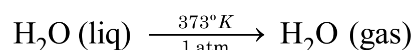
The first thing to appreciate is that when water is boiling the process is reversible. The slightest fall in temperature or increase in pressure will cause condensation.

NOTES

For a reversible process $\Delta G = 0$. That is to say, the conversion of water to steam or steam to water does not change the total free energy of one mole of water at 1 atm and 373ºK is equal to that of one mole of steam under the same conditions.

So, $\Delta G = \Delta H - T\Delta S = 0$ at the boiling point

Now clearly, for the change



$$H_2O\ (liq) \xrightarrow[1\ atm]{373°K} H_2O\ (gas)$$

there is a fixed ΔH and ΔS. From this it follows that there is a specific, unique temperature, T, at which boiling will occur.

You will see that below this temperature

$$\Delta H > T\Delta S$$

and that if water were to boil it would mean a change for which

$$\Delta G > 0$$

This means that boiling is impossible below this temperature.

Above 373°K

$$\Delta H < T\Delta S$$

And this means that for boiling now

$$\Delta G < 0.$$

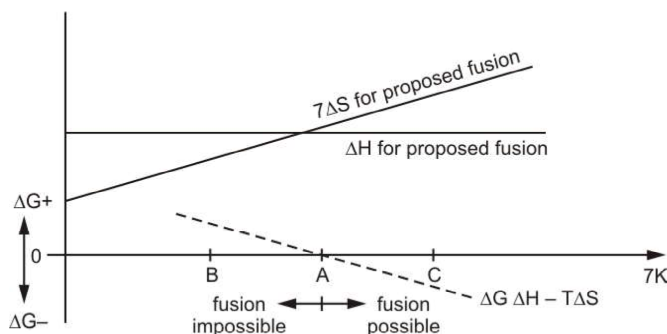


Fig. 2.4 *The Melting of a Solid*

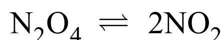
Figure 2.4 shows the thermodynamic treatment of a change of state, this time the melting of a solid. The horizontal line shows the enthalpy change for the fusion and this is taken to be independent of the temperature. The value of $T\Delta S$, on the other hand increases with the temperature. At A, $T\Delta S = \Delta H$ and this is the normal melting point. At B, $\Delta H > T\Delta S$, and melting is impossible because the change would give a positive ΔG. At C, melting can occur and the process here is irreversible,

$$\Delta G < 0$$

Free Energy and Chemical Equilibria

When dinitrogentetroxide N_2O_4 is warmed it will decompose to form nitrogen dioxide until an equilibrium is established.

NOTES



$$N_2O_4 \rightleftharpoons 2NO_2$$

It is now possible to relate the concentration by an equation

$$K = \frac{[NO_2]^2}{[N_2O_4]}$$

where K is the equilibrium constant. The value of K is fixed, but is temperature dependent.

We will now discuss chemical equilibria from a 'free-energy' point of view. How do they arise? The argument is that reaction fails to move in either direction because in both cases it would lead to an overall increase in free energy. At the equilibrium position, the system has a minimum free energy. This is shown in Figure 2.5.

Here the free energies of various N_2O_4/NO_2 mixtures are plotted against the compositions of the mixtures. When you study the graph, remember that it represents a fixed amount of matter, the chemical nature of which is changing.

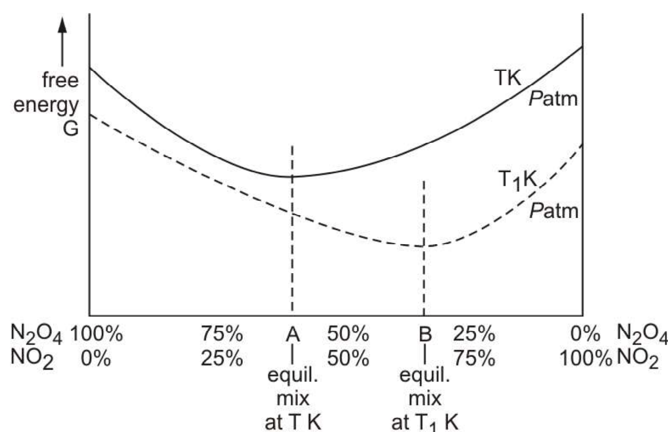


Fig. 2.5 *Free Energy and Chemical Equilibrium*

The graph shows that it's thermodynamically possible for dinitrogen tetroxide at TºK and p atm to form some nitrogen dioxide. Mixtures of N_2O_4 and NO_2 have lower free energies. Similarly, pure NO_2 at TºK and p atm will associate, because mixtures have a lower free energy than the pure chemical. The equilibrium mixture is shown at the lowest point of the graph; it is that mixture which has the lowest free energy at TºK and p atm. We can expect any mixture to the 'left' of the equilibrium to react.

It is well known that equilibrium constants change with temperature, i.e., the concentrations of the various chemicals at equilibrium change. This happens

because the free energies of reactants and products change as their temperature is changed;

$$G_{NO_2} = H_{NO_2} - T_{NO_2} \times S_{NO_2}$$

NOTES

When nitrogen dioxide or any other chemical is heated, its enthalphy, temperature, and entropy will change, so that its free energy will be affected. Also, the free energies of different chemicals are changed to different extents, so that the composition, of the mixture having a minimum free energy is now different. This is shown in Figure 2.3. The continuous line shows the free energy of mixtures at *TK*, while those at T_1K are shown by the broken line. *A* is the equilibrium mixture at *TK*, and *B* is that at T_1K.

How the Theory Explains the Facts

Since molecules of gases are relatively far apart, they may readily be compressed. The molecules are simply crowded closer together. It is more difficult to compress solid and liquid because of the close proximity of the molecules in the solid and liquid states.

The diffusion of gases can be explained by the motion of the molecules and the relatively large distance between the molecules. Since the molecules of each gas are free to move about among the molecules of one another, each gas is free to act relatively independently of other gases present in the mixture.

The explanation of the various gas laws.

Boyle's Law: When the volume of the gas is reduced at constant temperature, the more molecules are crowded together into a given span and collisions with the walls become more frequent, and the pressure is increased. Conversely, if the gas is expanded, fewer collisions take place, and the pressure is decreased.

Charle's Law: If the temperature of a gas is increased the velocity of the molecules and, subsequently, their kinetic energy is increased. With an increase in kinetic energy, the molecules of the gas move rapidly and number of collisions against the walls of the container increases. Now, if pressure is constant when the temperature is raised and walls of the container are elastic, like a rubber balloon, the walls expand to a great volume so that the number of collisions of the molecules per unit area of the walls (pressure) remain constant.

If the volume is constant, the pressure must be increased; thus explaining the increase in pressure of a gas with increase in temperature at constant volume.

Dalton's Law of Partial Pressures: In a mixture of gases, the molecules are free to move about and strike the walls of the container in an independent fashion and hence exert a pressure of its own. The total pressure of the gaseous mixture is thus, the sum of the pressures of all the molecular species present.

Graham's Law of Diffusion: The average kinetic energy of gas molecule is constant for a given temperature. The kinetic energy of a molecule is dependent on both mass and velocity, as given by the formula

NOTES

Average kinetic energy = 1/2 mv^2

where m is the mass of the molecule and v is its velocity. It follows from the equation that for less dense gas molecules, the speed is increased to maintain the constant kinetic energy. Similarly, the heavier gas molecules move more slowly than the heavier ones.

Liquifaction of Gases: The kinetic theory also explains the fact that gases may be liquefied by decreasing the temperature or increasing the pressure or both. In slowing down molecular movement by decreased temperature and forcing molecules nearer together by increased pressure, a stage is reached where intermolecular attraction some what offsets the kinetic energy of molecules. This causes the molecules coalesce and become a liquid.

Properties of Helmholtz Free Energy (A)

This function is defined as

$$A = E - TS$$

In order to understand the physical significance of A, consider the change in A when the system passes from state 1 to state 2, i.e.,

$$\Delta A = A_2 - A_1$$
$$= (E_2 - T_2S_2) - (E_1 - T_1S_1)$$
$$= \Delta E - (T_2S_2 - T_1S_1)$$

Under isothermal conditions, $T_2 = T_1 = T$, so that

$$\Delta A = E - T\,\Delta S \qquad ...(17)$$

Further, by definition of entropy, $\Delta S = \frac{q_r}{T}$

or $q_r = TdS$, so that we have

$$\Delta A = \Delta E - q_r \qquad ...(18)$$

But since from first law of thermodynamics, for a reversible isothermal process, $\Delta E - q_r = -W_m$, we have

$$\Delta A = -W_m \qquad ...(19)$$

Hence, at constant temperature, the maximum work obtainable from a system is at the expense of a decrease in the Helmholtz free energy of the system. This is why A is sometimes called the 'work function' or the maximum work content of a system.

Further, if we completely differentiate the equation $A = E - TS$, we get

$$dA = dE - TdS - SdT \qquad ...(20)$$

But $TdS = q_r$ and $q_r = dE + PdV$, so that equation (9.20) becomes

$$dA = dE - SdT - dE - PdV$$
$$dA = -SdT - PdV \qquad ...(21)$$

At constant volume, $dV = 0$ and

$$\left(\frac{\partial A}{\partial T}\right)_V = -S \qquad ...(22)$$

At constant temperature, $dT = 0$ and

$$\left(\frac{\partial A}{\partial V}\right)_T = -P \qquad ...(23)$$

Another equation which shows the variation of A with T can be obtained as given below.

Differentiating the quantity $\frac{A}{T}$ with respect to T at constant V, we get

$$\left[\frac{\partial\left(\frac{A}{T}\right)}{\partial T}\right]_V = \frac{T\left(\frac{\partial A}{\partial T}\right)_V - A}{T^2}$$

$$= \frac{-(A + TS)}{T^2}$$

But $\quad A + TS = E$

Therefore, $$\left[\frac{\partial\left(\frac{A}{T}\right)}{\partial T}\right]_V = -\frac{E}{T^2} \qquad ...(24)$$

Isothermal Change in Work Function: For an isothermal change, $dT = 0$, and hence, Equation (21) yields

$$dA = -PdV$$

For one mole of a perfect gas, P may be replaced by $\frac{RT}{V}$, so that

$$dA = -RT\frac{dV}{V} \qquad ...(25)$$

For an appreciable process, the increase ΔA in the work function can be obtained by integration of Equation (25) between the limits of the initial state 1 and final state 2; thus

$$\int_{A_1}^{A_2} dA = -\int_{V_1}^{V_2} RT\frac{dV}{V}$$

$$A_2 - A_1 = -RT\, ln\, \frac{V_2}{V_1}$$

or $$\Delta A = RT\, ln\, \frac{V_1}{V_2} \qquad ...(26)$$

Properties of Gibbs Free Energy (G)

This function is defined by

$$G = E - TS + PV \qquad ...(27)$$

NOTES

This definition may be written in two alternative but equivalent forms which are frequently employed. First, since H it equivalent to $E + PV$, hence

$$G = H - TS \qquad ...(28)$$

Second, since $A = E - TS$, it follows that

$$G = A + PV \qquad ...(29)$$

For a process taking place at constant pressure, the change in free energy is given by

$$\Delta G = \Delta A + P\,\Delta V \qquad ...(30)$$

If, in addition, the temperature is constant, $\Delta A = -W_m$ as seen from Equation (19), so that equation (9.30) becomes

$$\Delta G = -(W_m - P\,\Delta V) \qquad ...(31)$$

The quantity W_m represents maximum work obtainable in the given change and includes all types of work, such as electrical or surface work, in addition to work of expansion. The latter is equal to $P\,\Delta V$ and so $W_m - PdV$ represents the reversible work, exclusive of work of expansion, that can be obtained from a given change in state. The quantity $W_m - PdV$ is referred to as *net work.*

Thus, $\quad -\Delta G$ = Net work

Three decrease in free energy at constant T and P is equal to the maximum net work available for the given change in state which the process accompanies.

Again, from complete differential of Equation (28) we get

$$dG = dH - TdS - SdT \qquad ...(32)$$

But $\quad H = E + PV$, which on differentiation becomes

$dH = dE + PdV + VdP$. Also $TdS = dE + PdV$.

Making use of these identities into Equation (32),

We get

$$dG = dE + PdV + VdP - dE - PdV - SdT$$

$$= -SdT + VdP \qquad ...(33)$$

At constant pressure, $dP = 0$ and

$$\left(\frac{\partial G}{\partial T}\right)_P = -S \qquad ...(34)$$

At constant temperature, $dT = 0$ and

$$\left(\frac{\partial G}{\partial T}\right)_T = V \qquad ...(35)$$

An alternate equation which shows the variation of G with T is obtained by differentiating the quantity G/T with respect to T at constant P, namely,

$$\left[\frac{\partial\left(\frac{G}{T}\right)}{\partial T}\right]_P = \frac{T\left(\frac{\partial G}{\partial T}\right)_P - G}{T^2}$$

$$= -\frac{(G + TS)}{T^2}$$

But $\qquad G + TS = H.$

$$\text{Therefore,} \left[\frac{\partial\left(\frac{G}{T}\right)}{\partial T}\right]_P = -\frac{H}{T^2} \qquad ...(36)$$

Isothermal Changes in Free Energy Function: At constant temperature, $dT = 0$ and Equation (33) becomes,

$$dG = VdP \qquad ...(37)$$

For 1 mole of a perfect gas, $PV = RT$

or $\qquad V = \frac{RT}{P}$

Substituting this into Equation (37), we get

$$dG = \frac{RT}{P}\,dP \qquad ...(38)$$

Integrating between limits, Equation (38) takes the form

$$\int_{G_1}^{G_2} dG = RT \int_{P_1}^{P_2} \frac{dP}{P} \qquad ...(39)$$

$$G_2 - G_1 = RT \ln \frac{P_2}{P_1} \qquad ...(40)$$

or $\qquad \Delta G = RT \ln \frac{P_2}{P_1} \qquad ...(41)$

Generally, the free energy of a gas is related to the standard free energy $G°$. This is defined as the free energy of one mole of the gas at one atmosphere pressure. Equation (41) then becomes

$$G - G° = RT \ln \frac{P}{1} = RT \ln P$$

or $\qquad G = G° + RT \ln P \qquad ...(42)$

NOTES

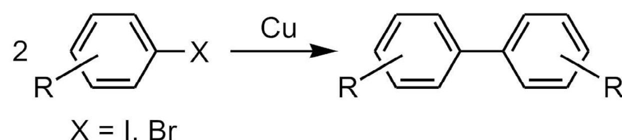
Check Your Progress

1. Define the first law of thermodynamics.
2. What is the need of second law of thermodynamics?
3. State the zeroth law of thermodynamics.
4. What do understand by third law of thermodynamics?
5. Elaborate on the bioenergetics.
6. Explain the coupling of chemical reactions.
7. Illustrate the Ullmann reaction.
8. Define the redox potential.
9. What is NADPH?
10. What does changes in free energy do?
11. What is Gibbs free energy?

2.9 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. According to the first law, the total energy of an isolated system remains constant during a specified change of state.
2. The second law of thermodynamics is needed because the first law of thermodynamics does not define the energy conversion process completely.
3. The zeroth law of thermodynamics defines thermal equilibrium and forms a basis for the definition of temperature: If two systems are each in thermal equilibrium with a third system, then they are in thermal equilibrium with each other.
4. The third law of thermodynamics states that a system's entropy approaches a constant value as the temperature approaches absolute zero. With the exception of non-crystalline solids (glasses) the entropy of a system at absolute zero is typically close to zero.
5. Bioenergetics is a field in biochemistry and cell biology that concerns energy flow through living systems. This is an active area of biological research that includes the study of the transformation of energy in living organisms and the study of thousands of different cellular processes, such as cellular respiration and the many other metabolic and enzymatic processes that lead to production and utilisation of energy in forms, such as Adenosine Triphosphate (ATP) molecules.

6. A coupling reaction in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst. In one important reaction type, a main group organometallic compound of the type R-M (R = organic fragment, M = main group centre) reacts with an organic halide of the type R'-X with formation of a new carbon-carbon bond in the product R-R'.
7. Coupling reactions are illustrated by the famous Ullmann reaction:



8. Redox potential also known as Oxidation Reduction Potential (ORP), the calculate value of redox potentials equal to the pe, E_0', or Eh is a measure of the tendency of a chemical species to acquire electrons from or lose electrons to an electrode and in this manner be reduced or oxidised respectively. Redox potential is measured in Volts (V), or milli Volts (mV).
9. NADPH is the reduced form of Nicotinamide Adenine Dinucleotide Phosphate ($NADP^+$). $NADP^+$ differs from Nicotinamide Adenine Dinucleotide (NAD^+) by the presence of an additional phosphate group on the 2' position of the ribose ring that carries the adenine moiety. This extra phosphate is added by NAD^+ kinase and removed by $NADP^+$ phosphatase.
10. The changes in free energy, DF or DG, are useful in determining the direction of spontaneous change and evaluating the maximum work that can be obtained from thermodynamic processes involving chemical or other types of reactions.
11. Gibbs free energy is a measure of the potential for reversible or maximum work that may be done by a system at constant temperature and pressure.

2.10 SUMMARY

- The first law states that energy of one form can be converted into an equivalent amount of energy of another form. But it does not tell that heat energy cannot be completely converted into an equivalent amount of work. There is thus need for another law, i.e., the second law of thermodynamics.
- The second law of thermodynamics helps us to determine the direction in which energy can be transferred. It also helps us to predict whether a given process or a chemical reaction can occur spontaneously, that is, its own accord. It also helps us to know the equilibrium conditions. The law is therefore, of great importance in chemistry.

NOTES

- The statement due to Clausius is it is impossible to construct a machine, which is able to convey heat by a cyclic process from one reservoir at a lower temperature to another at higher temperature unless work is done on the machine by some outside agency.
- The laws of thermodynamics define a group of physical quantities, such as temperature, energy, and entropy that characterise thermodynamic systems in thermodynamic equilibrium. The laws also use various parameters for thermodynamic processes, such as thermodynamic work and heat, and establish relationships between them.
- The second law of thermodynamics states that in a natural thermodynamic process, the sum of the entropies of the interacting thermodynamic systems never decreases. Another form of the statement is that heat does not spontaneously pass from a colder body to a warmer body.
- The zeroth law of thermodynamics provides for the foundation of temperature as an empirical parameter in thermodynamic systems and establishes the transitive relation between the temperatures of multiple bodies in thermal equilibrium.
- Bioenergetics is a field in biochemistry and cell biology that concerns energy flow through living systems. This is an active area of biological research that includes the study of the transformation of energy in living organisms and the study of thousands of different cellular processes, such as cellular respiration and the many other metabolic and enzymatic processes that lead to production and utilisation of energy in forms, such as Adenosine Triphosphate (ATP) molecules.
- Bioenergetics is the part of biochemistry concerned with the energy involved in making and breaking of chemical bonds in the molecules found in biological organisms. It can also be defined as the study of energy relationships and energy transformations and transductions in living organisms.
- Living organisms produce ATP from energy sources, mostly sunlight or O_2, mainly via oxidative phosphorylation. The terminal phosphate bonds of ATP are relatively weak compared with the stronger bonds formed when ATP is hydrolysed (broken down by water) to adenosine diphosphate and inorganic phosphate.
- A coupling reaction in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst. In one important reaction type, a main group organometallic compound of the type R-M (R = Organic Fragment, M = Main Group Centre) reacts with an organic halide of the type R'-X with formation of a new carbon-carbon bond in the product R-R'. The most common type of coupling reaction is the cross coupling reaction.

NOTES

- Heterocouplings: Heterocouplings combine two different partners, such as in the Heck reaction of an alkene (RC=CH) and an alkyl halide (R'-X) to give a substituted alkene. Heterocouplings are called cross-couplings.
- In aqueous solutions, redox potential is a measure of tendency of the solution to either gain or lose electrons when it is subjected to change by introduction of a new species. A solution with a higher (more positive) reduction potential than the new species will have a tendency to gain electrons from the new species (i.e. to be reduced by oxidising the new species) and a solution with a lower (more negative) reduction potential will have a tendency to lose electrons to the new species (i.e. to be oxidised by reducing the new species).
- Nicotinamide Adenine Dinucleotide (NAD) is a coenzyme central to metabolism. Found in all living cells, NAD is called a dinucleotide because it consists of two nucleotides joined through their phosphate groups. One nucleotide contains an adenine nucleobase and the other nicotinamide. NAD exists in two forms: an oxidised and reduced form, abbreviated as NAD^+ and NADH (H for hydrogen) respectively.
- In metabolism, nicotinamide adenine dinucleotide is involved in redox reactions, carrying electrons from one reaction to another. The cofactor is, therefore, found in two forms in cells: NAD^+ is an oxidising agent it accepts electrons from other molecules and becomes reduced. This reaction forms NADH, which can then be used as a reducing agent to donate electrons.
- NADPH is the reduced form of $NADP^+$. $NADP^+$ differs from NAD^+ by the presence of an additional phosphate group on the 2' position of the ribose ring that carries the adenine moiety. This extra phosphate is added by NAD^+ kinase and removed by $NADP^+$ phosphatase.
- A coupling reaction in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst. In one important reaction type, a main group organometallic compound of the type R-M (R = organic fragment, M = main group centre) reacts with an organic halide of the type R'-X with formation of a new carbon-carbon bond in the product R-R'.
- Free energy, in thermodynamics, energy-like property or state function of a system in thermodynamic equilibrium. Free energy has the dimensions of energy, and its value is determined by the state of the system and not by its history.
- Free energy is used to determine how systems change and how much work they can produce. It is expressed in two forms, i.e., the Helmholtz free energy F, sometimes called the work function and the Gibbs free energy G.

2.11 KEY WORDS

NOTES

- **Thermodynamics:** It is the branch of physical science that deals with the relationships between heat and other forms of energy.
- **First law of thermodynamics:** The first law of thermodynamics states that, when energy passes into or out of a system (as work, heat, or matter), the system's internal energy changes in accord with the law of conservation of energy.
- **Second law of thermodynamics:** The second law of thermodynamics states that in a natural thermodynamic process, the sum of the entropies of the interacting thermodynamic systems never decreases. Another form of the statement is that heat does not spontaneously pass from a colder body to a warmer body.
- **Bioenergetics:** Bioenergetics is a field in biochemistry and cell biology that concerns energy flow through living systems.
- **Coupling reaction:** A coupling reaction in organic chemistry is a general term for a variety of reactions where two fragments are joined together with the aid of a metal catalyst.
- **NADPH:** NADPH is the reduced form of $NADP^+$. $NADP^+$ differs from NAD^+ by the presence of an additional phosphate group on the 2' position of the ribose ring that carries the adenine moiety. This extra phosphate is added by NAD^+ kinase and removed by $NADP^+$ phosphatase.
- **Gibbs free energy:** It is a thermodynamic potential that can be utilised to determine the maximum of reversible work that may be completed by a thermodynamic system at a constant temperature and pressure.

2.12 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What are the failures of first law of thermodynamics?
2. What is the need of the second law of thermodynamics?
3. Elaborate on the first and second law of thermodynamics.
4. Interpret the principle of thermodynamics.
5. Give the applications of thermodynamics.
6. Define the term bioenergetics.
7. Illustrate the types of coupling of chemical reactions.
8. What is redox potential?

9. Explain about the NADP/NADPH.
10. Explain Gibbs free energy.
11. State the Helmholtz free energy.

NOTES

Long-Answer Questions

1. Briefly explain the first and second laws of thermodynamics giving appropriate examples.
2. Explain in detail about the principle and applications of thermodynamics.
3. Describe the concept of bioenergetics giving examples.
4. Analyse the coupling of chemical reactions with the help of various reactions.
5. Discuss briefly about the redox potential.
6. Describe the basic concept of NADP/NADPH.
7. Derive Gibbs-Helmholtz equation. What are its applications?
8. Briefly explain about the Gibbs and Helmholtz free energy.

2.13 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 3 NATURAL RADIATIONS

NOTES

Structure

3.0 INTRODUCTION

Natural radiation originates from a variety of sources, both natural and artificial. These include both cosmic radiation and environmental radioactivity from naturally occurring radioactive materials (such as, Radon and Radium), as well as man-made medical X-rays, fallout from nuclear weapons testing and nuclear accidents. Light or visible light is electromagnetic radiation within the portion of the electromagnetic spectrum that can be perceived by the human eye. Light sometimes refers to electromagnetic radiation of any wavelength, whether visible or not. In this sense, Gamma rays, X-rays, Microwaves and Radio waves are also light.

Absorption of light is the process of light is absorbed and converted into energy. The absorption of light is therefore directly proportional to the frequency. Absorption depends on the electromagnetic frequency of the light and object's nature of atoms.

A quantum mechanical system or particle that is bound and confined spatially can only take on certain discrete values of energy, called energy levels. This contrasts with classical particles, which can have any amount of energy. The term is commonly used for the energy levels of the electrons in atoms, ions, or molecules, which are bound by the electric field of the nucleus, but can also refer to energy levels of nuclei vibrational or rotational energy levels in molecules.

NOTES

Spin is an intrinsic form of angular momentum carried by elementary particles, composite particles (hadrons), and atomic nuclei. Spin is one of two types of angular momentum in quantum mechanics, the other being orbital angular momentum. The orbital angular momentum operator is the quantum-mechanical counterpart to the classical angular momentum of orbital revolution and appears when there is periodic structure to its wave function as the angle varies.

In quantum mechanics, an excited state of a system (such as, an atom, molecule or nucleus) is any quantum state of the system that has a higher energy than the ground state (that is, more energy than the absolute minimum). The ground state of a quantum-mechanical system is its lowest-energy state; the energy of the ground state is known as the zero-point energy of the system.

A biomolecule or biological molecule is a used for the molecules present in organisms that are essential to one or more typically biological processes, such as cell division, morphogenesis, or development. Biomolecules include large macromolecules (or polyanions), such as Proteins, Carbohydrates, Lipids, and Nucleic Acids, as well as small molecules, such as primary metabolites, secondary metabolites and natural products. A more general name for this class of material is biological materials.

In this unit, you will study about the natural radiations, properties of light, absorption of light, energy state of atom, spin property of electron, ground state and excited state of atoms, bio-molecules and their effects.

3.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the natural radiation
- Explain the properties of light
- Discuss about the absorption of light
- Analyse the energy state of atom
- Elaborate on the spin property of electron
- Interpret the ground state and excited state of atoms
- Define the bio-molecules and their effects

3.2 NATURAL RADIATIONS

Natural radiation is a measure of the level of ionising radiation present in the environment at a particular location which is not due to deliberate introduction of

NOTES

radiation sources. Natural radiation originates from a variety of sources, both natural and artificial. These include both cosmic radiation and environmental radioactivity from naturally occurring radioactive materials (such as radon and radium), as well as man-made medical X-rays, fallout from nuclear weapons testing and nuclear accidents.

Natural Radiations are:

- Cosmic Radiation
- Terrestrial Radiation

3.2.1 Cosmic Radiation

Cosmic radiation is an ionising radiation produced when primary photons and α-particles from outside the solar system interact with components of the earth's atmosphere. A second source of cosmic radiation is the release of charged particles from the sun, which become significant during periods of solar flare ('Sun Storm'). Ionising radiation is a natural part of the environment in which we live, present in the earth, buildings, the food we eat, and even in the bones of our bodies. The other type is non-ionising radiation which includes Ultra Violet (UV) light, radio waves, and microwaves. Humans, animals, and plants have all evolved in an environment with a background of natural radiation and, with few exceptions, it is not a significant risk to health.

The amount of cosmic radiation that reaches the earth from the sun and outer space varies: its energy is effectively absorbed by the atmosphere and is also affected by the earth's magnetic field. The effect on the body will depend on the latitude and altitude at which the individual is exposed, and on the length of time of exposure.

Cosmic background radiation is electromagnetic radiation from the ***Big Bang***. The origin of this radiation depends on the region of the spectrum that is observed. One component is the cosmic microwave background. This component is redshifted photons that have freely streamed from an epoch when the Universe became transparent for the first time to radiation. Its discovery and detailed observations of its properties are considered one of the major confirmations of the Big Bang. The discovery of the cosmic background radiation suggests that the early universe was dominated by a radiation field, a field of extremely high temperature and pressure.

NOTES

Cosmic radiation may be measured directly using sophisticated instruments, as was done routinely in the Concorde supersonic transport, or it can be estimated using a computer program integrating the route, time at each altitude, and phase of the solar cycle to calculate the radiation dose for any given flight. Several research organisations have confirmed actual measurements taken on board an aircraft to verify the computer estimations.

3.2.2 Terrestrial Radiation

The portion of the natural background radiation that is emitted by naturally occurring radioactive materials, such as uranium, thorium, and radon in the earth. Terrestrial radiation refers to sources of radiation that are in the soil, water, and vegetation. The major isotopes of concern for terrestrial radiation are Potassium, Uranium and the decay products of Uranium, such as Thorium, Radium, and Radon. Terrestrial radiation includes an external exposure caused by these radionuclides. Terrestrial radiation, for the purpose of shown in Table 3.1, only includes sources that remain external to the body. The major radionuclides of concern are Potassium, Uranium and Thorium and their decay products, some of which, like Radium and Radon are intensely radioactive but occur in low concentrations. Most of these sources have been decreasing, due to radioactive decay since the formation of the Earth, because there is no significant amount currently transported to the Earth. Thus, the present activity on earth from Uranium-238 is only half as much as it originally was because of its 4.5 billion year half-life, and Potassium-40 (half-life 1.25 billion years) is only at about 8% of original activity. But during the time that humans have existed the amount of radiation has decreased very little.

Many shorter half-life (and thus more intensely radioactive) isotopes have not decayed out of the terrestrial environment because of their on-going natural production. Examples of these are Radium-226 (decay product of Thorium-230 in decay chain of Uranium-238) and Radon-222 (a decay product of Radium-226 in said chain).

Thorium and Uranium (and their daughters) primarily undergo α and β– decay, and are not easily detectable. However, many of their daughter products are strong γ–emitters. Thorium-232 is detectable via a 239 k eV peak from lead-212, 511, 583 and 2614 ke V from Thallium-208, and 911 and 969 k eV from Actinium-228. Uranium-238 manifests as 609, 1120, and 1764 k eV peaks of Bismuth-214 (cf. the same peak for atmospheric radon). Potassium-40 is detectable directly via its 1461 k eV γ–peak.

Table 3.1 *Average Annual Human Exposure to Ionizing Radiation in Milli Sieverts (MSV) Per Year*

NOTES

Radiation source	World	US	Japan	Remark
Inhalation of Air	1.26	2.28	0.40	Mainly from Radon, Depends on Indoor Accumulation
Ingestion of Food & Water	0.29	0.28	0.40	(K-40, C-14, etc.)
Terrestrial Radiation from Ground	0.48	0.21	0.40	Depends on Soil and Building Material
Cosmic Radiation from Space	0.39	0.33	0.30	Depends on Altitude
Subtotal (natural)	2.40	3.10	1.50	Sizeable Population Groups Receive 10–20 mSv
Medical	0.60	3.00	2.30	Worldwide Figure Excludes Radiotherapy; US Figure is Mostly CT Scans and Nuclear Medicine.
Consumer Items	–	0.13		Cigarettes, Air Travel, Building Materials, etc.
Atmospheric Nuclear Testing	0.005	–	0.01	Peak Of 0.11 mSv In 1963 And Declining Since; Higher Near Sites
Occupational Exposure	0.005	0.005	0.01	Worldwide Average to Workers only is 0.7 mSv, Mostly Due to Radon in Mines; US is Mostly Due to Medical and Aviation Workers.
Chernobyl Accident	0.002	–	0.01	Peak of 0.04 mSv in 1986 and Declining since; Higher Near Site
Nuclear Fuel Cycle	0.0002		0.001	up to 0.02 mSv Near Sites; Excludes Occupational Exposure
Other	–	0.003		Industrial, Security, Medical, Educational, and Research
Subtotal (artificial)	0.61	3.14	2.33	
Total	3.01	6.24	3.83	**Milli Sieverts per Year**

3.3 PROPERTIES OF LIGHT

Light is a form of energy which is given by luminous objects. The Sun, bulb, candle, etc., are luminous objects. Other objects which do not give out light are called non-luminous objects. Light can pass through transparent materials. Glass, water, clear plastic and air are transparent materials. Light cannot pass through translucent materials .A tracing paper, frosted glass and waxed paper are translucent materials.

Light or visible light is electromagnetic radiation within the portion of the electromagnetic spectrum that is perceived by the human eye. Visible light is usually defined as having wavelengths in the range of 400–700 nanometres (nm), between the Infrared (IR) (with longer wavelengths) and the ultra Violet (with shorter wavelengths). This wavelength means a frequency range of roughly 430–750 Terahertz (THz).

NOTES

The primary properties of visible light are intensity, propagation-direction, frequency or wavelength spectrum and polarisation. Its speed in a vacuum, 299 792 458 metres per second (m/s), is one of the fundamental constants of nature, as with all types of Electromagnetic Radiation (EMR), light is found in experimental conditions to always move at this speed in a vacuum.

There are 7 basic properties of light:

- Reflection of Light
- Refraction of Light
- Diffraction of Light
- Interference of Light
- Polarisation of Light
- Dispersion of Light
- Scattering of Light

Reflection of Light: Reflection is the change in direction of a wave front at an interface between two different media so that the wave front returns into the medium from which it originated. The common examples include the reflection of light, sound and water waves. The law of reflection says that for specular reflection the angle at which the wave is incident on the surface equals the angle at which it is reflected. Mirrors exhibit specular reflection.

In acoustics, reflection causes echoes and is used in sonar. In geology, it is important in the study of seismic waves. Reflection is observed with surface waves in bodies of water. Reflection is observed with many types of electromagnetic wave, besides visible light. Reflection of Very High Frequency (VHF) and higher frequencies is important for radio transmission and for radar. Even hard X-rays and gamma rays can be reflected at shallow angles with special 'Grazing' mirrors.

Refraction of Light: In physics, refraction is the change in direction of a wave passing from one medium to another or from a gradual change in the medium. Refraction of light is the most commonly observed phenomenon, but other waves, such as sound waves and water waves also experience refraction. How much a wave is refracted is determined by the change in wave speed and the initial direction of wave propagation relative to the direction of change in speed. For light, refraction follows Snell's law, which states that, for a given pair of media, the ratio of the sines of the angle of incidence θ_1 and angle of refraction θ_2 is equal to the ratio of phase velocities (v_1/v_2) in the two media, or equivalently, to the indices of refraction (n_2/n_1) of the two media.

$$\frac{\sin\theta_1}{\sin\theta_2} = \frac{v_i}{v_2} = \frac{n_2}{n_1}$$

NOTES

Optical prisms and lenses use refraction to redirect light, as does the human eye. The refractive index of materials varies with the wavelength of light, and thus the angle of the refraction also varies correspondingly. This is called dispersion and causes prisms and rainbows to divide white light into its constituent spectral colours.

Diffraction of Light: Diffraction refers to various phenomena that occur when a wave encounters an obstacle or opening. It is defined as the bending of waves around the corners of an obstacle or through an aperture into the region of geometrical shadow of the obstacle/aperture. The diffracting object or aperture effectively becomes a secondary source of the propagating wave. Italian scientist Francesco Maria Grimaldi coined the word diffraction and was the first to record accurate observations of the phenomenon in 1660.

These effects also occur when a light wave travels through a medium with a varying refractive index, or when a sound wave travels through a medium with varying acoustic impedance all waves diffract, including gravitational waves, water waves, and other electromagnetic waves, such as X-rays and radio waves. Furthermore, quantum mechanics also demonstrates that matter possesses wave-like properties, and hence, undergoes diffraction (which is measurable at subatomic to molecular levels).

Interference of Light: In physics, interference is a phenomenon in which two waves superpose to form a resultant wave of greater, lower, or the same amplitude. Constructive and destructive interference result from the interaction of waves that are correlated or coherent with each other, either because they come from the same source or because they have the same or nearly the same frequency. Interference effects can be observed with all types of waves, for example, light, radio, acoustic, surface water waves, gravity waves, or matter waves. The resulting images or graphs are called interferograms.

The principle of superposition of waves states that when two or more propagating waves of the same type are incident on the same point, the resultant amplitude at that point is equal to the vector sum of the amplitudes of the individual waves. If a crest of a wave meets a crest of another wave of the same frequency at the same point, then the amplitude is the sum of the individual amplitudes this is constructive interference. If a crest of one wave meets a trough of another wave, then the amplitude is equal to the difference in the individual amplitudes this is known as ***destructive interference***.

Constructive interference occurs when the phase difference between the waves is an even multiple of π (180°), whereas destructive interference occurs when the difference is an odd multiple of π. If the difference between the phases is intermediate between these two extremes, then the magnitude of the displacement of the summed waves lies between the minimum and maximum values.

NOTES

Polarisation of Light: Polarisation is a property applying to transverse waves that specifies the geometrical orientation of the oscillations. In a transverse wave, the direction of the oscillation is perpendicular to the direction of motion of the wave. A simple example of a polarised transverse wave is vibrations traveling along a taut string, for example in a musical instrument like a guitar string. Depending on how the string is plucked, the vibrations can be in a vertical direction, horizontal direction, or at any angle perpendicular to the string. In contrast, in longitudinal waves, such as sound waves in a liquid or gas, the displacement of the particles in the oscillation is always in the direction of propagation, so these waves do not exhibit polarisation. Transverse waves that exhibit polarisation include electromagnetic, such as light and radio waves, gravitational waves, and transverse sound waves (shear waves) in solids.

Dispersion of Light: In optics, dispersion is the phenomenon in which the phase velocity of a wave depends on its frequency. Media having this common property may be termed dispersive media. Sometimes the term chromatic dispersion is used for specificity. Although the term is used in the field of optics to describe light and other electromagnetic waves, dispersion in the same sense can apply to any sort of wave motion, such as acoustic dispersion in the case of sound and seismic waves, in gravity waves (ocean waves), and for telecommunication signals along transmission lines (such as coaxial cable) or optical fiber. Physically, dispersion translates in a loss of kinetic energy through absorption.

Scattering of Light: Scattering is a term used in physics to describe a wide range of physical processes where moving particles or radiation of some form, such as light or sound, is forced to deviate from a straight trajectory by localized non-uniformities (including particles and radiation) in the medium through which they pass. In conventional use, this also includes deviation of reflected radiation from the angle predicted by the law of reflection. Reflections of radiation that undergo scattering are often called diffuse reflections and unscattered reflections are called ‘Specular Reflections’ (mirror-like).

3.4 ABSORPTION OF LIGHT

Light absorption is a process by which light is absorbed and converted into energy. An example of this process is photosynthesis in plants. However, light absorption does not happen absolutely in plants, but in all creatures/inorganic substances. Absorption depends on the electromagnetic frequency of the light and object’s nature of atoms. The absorption of light is therefore directly proportional to the frequency. If they are complementary, light is absorbed. If they are not complementary, then the light passes through the object or gets reflected. These processes usually occur at the same time because the light is usually transmitted at various frequencies. (For instance, sunlight also comprises lights of various

NOTES

frequencies; from around 400 to 800 nm). Therefore, most objects selectively absorb, transmit, or reflect the light. When light is absorbed heat is generated. So the selective absorption of light by a particular material occurs because the frequency of the light wave matches the frequency at which electrons in the atoms of that material vibrate.

Absorption depends on the state of an object's electron. All electrons vibrate at a specific frequency, which is known as their 'Natural' frequency. When light interacts with an atom of the same frequency, the electrons of the atom become excited and start vibrating. During this vibration, the electrons of the atom interact with neighbouring atoms and convert this vibrational energy into thermal energy. Consequently, the light energy is not to be seen again, that is why absorption differs from reflection and transmission. And since different atoms and molecules have different natural frequencies of vibration, they selectively absorb different frequencies of visible light. For example, organic molecules are good at absorbing light. If an organic molecule has electrons that have a high natural frequency then they absorb the light which has a high frequency as well. The longer the conjugated system (conjugated system is a system of connected π-orbitals with delocalized electrons), the longer the wavelength of the light absorbed.

By Relying on this method, physicists are able to determine and identify the properties and material composition of an object by observing which frequencies of light it absorbs. While some materials are opaque to some wavelengths of light, they are transparent to others. Wood, for example is opaque to all forms of visible light. Glass and water on the other hand are opaque to ultraviolet light, but transparent to visible light.

Light Absorption and Colours

Absorption of electromagnetic radiation requires an opposite-field, i.e., the field which has the opposite coefficient in the same mode. A good, for example of this is colour. If a material or matter absorbs light of certain wavelengths (or colours) of the spectrum, an observer will not see these colours in the reflected light. On the other hand, if certain wavelengths of colours are reflected from the material, these are the colours that the observer will see. For example, leaves contain the pigment *Chlorophyll*, which absorbs the blue and red colours of the spectrum and reflects green therefore leaves appear green. To the naked eye, reflected light often appears to be refracted into several colours of the spectrum. As a result, light absorption is related to matter's frequency (and frequency of light also) and wavelength of light.

Quantifying Absorption

Many approaches can potentially quantify radiation absorption, with key examples following.

- The absorption coefficient along with some closely related derived quantities.
- The attenuation coefficient (NB used infrequently with meaning synonymous with 'Absorption Coefficient').

- The Molar attenuation coefficient (also called 'Molar Absorptivity'), which is the absorption coefficient divided by molarity.
- The mass attenuation coefficient (also called 'Mass Extinction Coefficient'), which is the absorption coefficient divided by density.
- The absorption cross section and scattering cross-section, related closely to the absorption and attenuation coefficients, respectively.
- '*Extinction*' in astronomy, which is equivalent to the attenuation coefficient.
- Other measures of radiation absorption, including penetration depth and skin effect, propagation constant, attenuation constant, phase constant, and complex wavenumber, complex refractive index and extinction coefficient, complex dielectric constant, electrical resistivity and conductivity.
- Related measures, including absorbance (also called 'Optical Density') and optical depth (also called 'Optical Thickness')

All these quantities measure, at least to some extent, how well a medium absorbs radiation. Which among them practitioners use varies by field and technique, often due simply to the convention.

Applications

Understanding and measuring the absorption of electromagnetic radiation has a variety of applications.

- In radio propagation, it is represented in non-line-of-sight propagation. For example, computation of radio wave attenuation in the atmosphere used in satellite link design.
- In meteorology and climatology, global and local temperatures depend in part on the absorption of radiation by atmospheric gases (such as, in the greenhouse effect) and land and ocean surfaces.
- In medicine, X-rays are absorbed to different extents by different tissues (bone in particular), which is the basis for X-ray imaging.
- In chemistry and materials science, different materials and molecules absorb radiation to different extents at different frequencies, which allows for material identification.
- In optics, sunglasses, coloured filters, dyes, and other such materials are designed specifically with respect to which visible wavelengths they absorb, and in what proportions they are in.
- In biology, photosynthetic organisms require that light of the appropriate wavelengths be absorbed within the active area of chloroplasts, so that the light energy can be converted into chemical energy within sugars and other molecules.

NOTES

- In physics, the D-region of Earth's ionosphere is known to *significantly absorb radio signals* that fall within the high-frequency electromagnetic spectrum.
- In nuclear physics, absorption of nuclear radiations can be used for measuring the fluid levels, densitometry or thickness measurements.

3.5 ENERGY STATE OF ATOMS

A quantum mechanical system or particle that is bound it means that, confined spatially can only take on certain discrete values of energy, called energy state of atom. This contrasts with classical particles, which can have any amount of energy. The term is commonly used for the energy levels of the electrons in atoms, ions, or molecules, which are bound by the electric field of the nucleus, but can also refer to energy levels of nuclei or vibrational or rotational energy levels in molecules. The energy spectrum of a system with such discrete energy levels is said to be quantized.

In chemistry and atomic physics, an electron shell, or principal energy level, may be thought of as the orbit of one or more electrons around an atom's nucleus. The closest shell to the nucleus is called the '1 Shell' (also called 'K Shell'), followed by the '2 Shell' (or 'L Shell'), then the '3 Shell' (or 'M Shell'), and so on farther and farther from the nucleus. The shells correspond with the principal quantum numbers ($n = 1, 2, 3, 4$ ...) or are labelled alphabetically with letters used in the X-ray notation (K, L, M, N…).

Each shell can contain only a fixed number of electrons: The first shell can hold up to two electrons, the second shell can hold up to eight (2 + 6) electrons, the third shell can hold up to 18 (2 + 6 + 10) and so on. The general formula is that the nth shell can in principle hold up to $2(n^2)$ electrons. Since electrons are electrically attracted to the nucleus, an atom's electrons will generally occupy outer shells only if the more inner shells have already been completely filled by other electrons.

If the potential energy is set to zero at infinite distance from the atomic nucleus or molecule, the usual convention, then bound electron states have negative potential energy.

If an atom, ion, or molecule is at the lowest possible energy level, it and its electrons are said to be in the ground state. If it is at a higher energy level, it is said to be excited state, or any electrons that have higher energy than the ground state are excited. If more than one quantum mechanical state is at the same energy, the energy levels are 'Degenerate'. They are then called 'Degenerate Energy' levels.

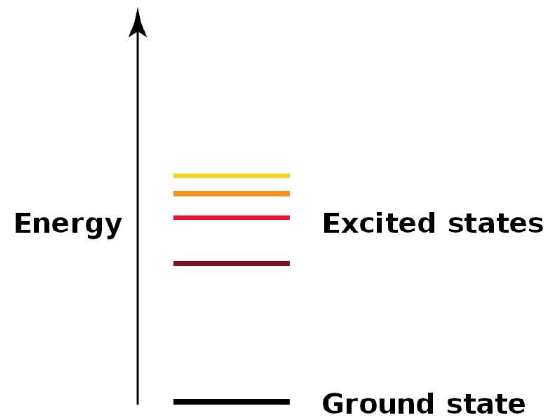


Fig 3.1 *Energy Levels for an Electron in an Atom*

NOTES

Intrinsic Energy Levels of Atom

In the formulas for energy of electrons at various levels given below in an atom, the zero point for energy is set when the electron in question has completely left the atom, i.e., when the electron's principal quantum number *n* = ". When the electron is bound to the atom in any closer value of *n,* the electron's energy is lower and is considered negative.

Orbital state energy level: atom/ion with nucleus + one electron

Assume there is one electron in a given atomic orbital in a hydrogen-like atom (ion). The energy of its state is mainly determined by the electrostatic interaction of the (negative) electron with the (positive) nucleus. The energy levels of an electron around a nucleus are given by:

$$E_n = -hcR_\infty \frac{Z^2}{n^2}$$

Where *R*„ is the Rydberg Constant

Z is the Atomic Number

n is the principal quantum number

h is Planck's Constant

c is the Speed of Light.

For hydrogen-like atoms (ions) only, the Rydberg levels depend only on the principal quantum number *n*.

This equation is obtained from combining the Rydberg formula for any hydrogen-like element with E = hν = hc/λ assuming that the principal quantum number n above = n_1 in the Rydberg formula and $n_2 = \infty$ (principal quantum number of the energy level the electron descends from, when emitting a photon). The Rydberg formula was derived from empirical spectroscopic emission data.

$$\frac{1}{\lambda} = RZ^2 \left(\frac{1}{n_1^2} - \frac{1}{n_2^2} \right)$$

NOTES

An equivalent formula can be derived quantum mechanically from the time-independent Schrödinger equation with a kinetic energy Hamiltonian operator using a wave function as an Eigen function to obtain the energy levels as eigenvalues, but the Rydberg constant would be replaced by other fundamental physics constants.

Electron-Electron Interactions in Atoms

If there is more than one electron around the atom, electron-electron-interactions raise the energy level. These interactions are often neglected if the spatial overlap of the electron wave functions is low.

For multi-electron atoms, interactions between electrons cause the preceding equation to be no longer accurate as stated simply with Z as the atomic number. A simple way to understand this is as a *Shielding Effect*, where the outer electrons see an effective nucleus of reduced charge, since the inner electrons are bound tightly to the nucleus and partially cancel its charge. This leads to an approximate correction where Z is substituted with an *Effective Nuclear Charge* (ENC) symbolized as Z_{eff} that depends strongly on the principal quantum number.

$$E_{n,l} = -hcR_\infty \frac{{Z_{\text{eff}}}^2}{n^2}$$

In such cases, the orbital types (determined by the Azimuthal Quantum Number l) as well as their levels within the molecule affect Z_{eff} and therefore also affect the various atomic electron energy levels. The Aufbau principle of filling an atom with electrons for an electron configuration takes these differing energy levels into account. For filling an atom with electrons in the ground state, the lowest energy levels are filled first and consistent with the *Pauli Exclusion Principle,* the *Aufbau principle,* and *Hund's rule.*

3.6 SPIN PROPERTIES OF ELECTRONS

Spin is an intrinsic form of angular momentum carried by elementary particles, composite particles (hadrons), and atomic nuclei. Spin is one of two types of angular momentum in quantum mechanics, the other being orbital angular momentum. The orbital angular momentum operator is the quantum-mechanical counterpart to the classical angular momentum of orbital revolution and appears when there is periodic structure to its wave function as the angle varies. For photons, spin is the quantum-mechanical counterpart of the polarisation of light; for electrons, the spin has no classical counterpart.

NOTES

The existence of electron spin angular momentum is inferred from experiments, such as the Stern–Gerlach experiment, in which silver atoms were observed to possess two possible discrete angular momenta despite having no orbital angular momentum. The existence of the electron spin can also be inferred theoretically from spin–statistics theorem and from the Pauli Exclusion Principle and vice versa, given the particular spin of the electron, one may derive the Pauli Exclusion Principle.

Spin is described mathematically as a vector for some particles, such as photons, and as spinors and bispinors for other particles, such as electrons. Spinors and bispinors behave similarly to vectors: they have definite magnitudes and change under rotations; however, they use an unconventional Direction'. All elementary particles of a given kind have the same magnitude of spin angular momentum, though its direction may change. These are indicated by assigning the particle a spin quantum number.

The System International (SI) unit of spin is the same as classical angular momentum (i.e. N·m·s or $kg·m^2·s^{-1}$). In practice, spin is given as a dimensionless spin quantum number by dividing the spin angular momentum by the reduced Planck constant ‘, which has the same dimensions as angular momentum, although this is not the full computation of this value. Very often, the ‘Spin Quantum Number’ is simply called ‘Spin’. The fact that it is a quantum number is implicit.

Quantum Number of Spin Electron

In atomic physics, the spin quantum number is a quantum number (designated m_s) which describes the intrinsic angular momentum (or spin angular momentum, or simply spin) of an electron or other particle. The phrase was originally used to describe the fourth of a set of quantum numbers (the principal quantum number n, the azimuthal quantum number l, the magnetic quantum number m, and the spin quantum number m_s), which completely describe the quantum state of an electron. The name comes from a physical spinning of the electron about an axis that was proposed by Uhlenbeck and Goudsmit, and the value of m_s is the component of spin angular momentum parallel to a given direction (the z–axis), which can be either +1/2 or –1/2 (in units of the reduced Planck constant).

As a solution for a certain partial differential equation, the quantized angular momentum can be written as:

$$\|s\| = \sqrt{s(s+1)}\hbar$$

Where

s is the quantized spin vector

$\|s\|$ is the norm of the spin vector

s is the spin quantum number associated with the spin angular momentum

$\hbar$ is the reduced Planck constant.

3.7 GROUND STATE AND EXCITED STATE OF ATOMS

NOTES

The **ground state** of a quantum-mechanical system is its lowest-energy state; the energy of the ground state is known as the zero-point energy of the system. While molecule absorbed energy it will get exited position. An excited state is any state with energy greater than the ground state. In quantum field theory, the ground state is usually called the vacuum state or the vacuum. If more than one ground state exists, they are said to be degenerate. Many systems have degenerate ground states. Degeneracy occurs whenever there exists a unitary operator that acts non-trivially on a ground state and commutes with the Hamiltonian of the system.

For example according to the third law of thermodynamics, a system at absolute zero temperature exists in its ground state; thus, its entropy is determined by the degeneracy of the ground state. Many systems, such as a perfect crystal lattice, have a unique ground state and therefore have zero entropy at absolute zero. It is also possible for the highest excited state to have absolute zero temperature for systems that exhibit negative temperature.

In quantum mechanics, an **excited state** of a system (such as an atom, molecule or nucleus) is any quantum state of the system that has a higher energy than the ground state (that is, more energy than the absolute minimum). Excitation is an elevation in energy level above an arbitrary baseline energy state. In physics there is a specific technical definition for energy level which is often associated with an atom being raised to an excited state. The temperature of a group of particles is indicative of the level of excitation (with the notable exception of systems that exhibit negative temperature). The lifetime of a system in an excited state is usually short: spontaneous or induced emission of a quantum of energy (such as a photon or a phonon) usually occurs shortly after the system is promoted to the excited state, returning the system to a state with lower energy (a less excited state or the ground state). This return to a lower energy level is often loosely described as decay and is the inverse of excitation.

Atomic Excitation

A simple example of this concept comes by considering the hydrogen atom.

The ground state of the hydrogen atom corresponds to having the atom's single electron in the lowest possible orbital (that is, the spherically symmetric '1s' wave function, which, so far, has demonstrated to have the lowest possible quantum numbers). By giving the atom additional energy (for example, by the absorption of a photon of an appropriate energy), the electron is able to move into an excited state (one with one or more quantum numbers greater than the minimum possible). If the photon has too much energy, the electron will cease to be bound to the atom, and the atom will become ionised. After excitation the atom may return to

NOTES

the ground state or a lower excited state, by emitting a photon with a characteristic energy. Emission of photons from atoms in various excited states leads to an electromagnetic spectrum showing a series of characteristic emission lines (including, in the case of the hydrogen atom, the Lyman, Balmer, Paschen and Brackett series.) An atom in a high excited state is termed a Rydberg atom. A system of highly excited atoms can form a long-lived condensed excited state, e.g., a condensed phase made completely of excited atoms: Rydberg matter. Hydrogen can also be excited by heat or electricity.

The excitation of a system (an atom or molecule) from one excited state to a higher energy excited state with the absorption of a photon is called Excited State Absorption (ESA). Excited state absorption is possible only when an electron has been already excited from the ground state to a lower excited state. The excited state absorption is usually an undesired effect, but it can be useful in up conversion pumping. Excited state absorption measurements are done using pump-probe techniques, such as flash photolysis. However, it is not easy to measure them compared to ground-state absorption and in some cases complete bleaching of the ground state is required to measure excited state absorption.

3.8 BIOMOLECULES AND THEIR EFFECTS

A biomolecule or biological molecule is a loosely used term for molecules present in organisms that are essential to one or more typically biological processes, such as cell division, morphogenesis, or development. Biomolecules include large macromolecules (or polyanions), such as proteins, carbohydrates, lipids, and nucleic acids, as well as small molecules, such as primary metabolites, secondary metabolites and natural products. A more general name for this class of material is biological materials. Biomolecules are an important element of living organisms, those biomolecules are often endogenous, produced within the organism but organisms usually need exogenous biomolecules, for example certain nutrients, to survive.

Biology and its subfields of biochemistry and molecular biology study biomolecules and their reactions. Most biomolecules are organic compounds, and just four elements oxygen, carbon, hydrogen, and nitrogen make up 96% of the human body's mass. But many other elements, such as the various biometals, are also present in small amounts.

The uniformity of both specific types of molecules (the biomolecules) and of certain metabolic pathways are invariant features among the wide diversity of life forms; thus these biomolecules and metabolic pathways are referred to as Biochemical Universals or '*Theory of Material Unity of the Living Beings*', a unifying concept in biology, along with cell theory and evolution theory.

Types of Biomolecules

A diverse range of biomolecules exist, including

Small molecules:

NOTES

- Lipids, fatty acids, glycolipids, sterols, monosaccharides
- Vitamins
- Hormones, neurotransmitters
- Metabolites

Table 3.2 *Monomers, Oligomers and Polymers*

Biomonomer	Bio-oligo	Biopolymers	Polymerisation process	Covalent bond name between monomers
Amino acids	Oligopeptides	Polypeptides, proteins (Hemoglobin)	Polycondensation	Peptide bond
Monosaccharides	Oligosaccharides	Polysaccharides (cellulose)	Polycondensation	Glycosidic bond
Isoprene	Terpenes	Polyterpenes: cis-1,4-polyisoprene natural rubber and trans-1,4-polyisoprene gutta-percha	Polyaddition	
Nucleotides	Oligonucleotides	Polynucleotides, nucle ic acids (DNA, RNA)		Phosphodiester bond

Effect of Biomolecules

- Biomolecules are known to play a crucial role in the growth and microstructure formation of some biological nanocomposites that exhibit superior mechanical properties. Thus, biomolecules are considered promising compounds to manipulate the structure of bio-inspired materials.
- Biomolecules are an organic molecule that includes carbohydrates, protein, lipids, and nucleic acids. They are important for the survival of living cells. Some of valuable biomolecules have huge demand, which cannot be fulfilled from their renewable resources. Microbes have been used as a cell factory for their alternative production.
- The physico-chemical viewpoint of using extreme conditions is to explore the effect of temperature and pressure on the conformation, the dynamics and the reactions of biomolecules. The unique properties of biomolecules are determined by the delicate balance between internal interactions in the biomolecules which compete with interactions with the solvent. The primary source of the dynamical behaviour of biomolecules is the free volume of the system and this may be expected to decrease with increasing pressure. As temperature effects act via an increased kinetic energy as well as free volume, it follows that the study of the combined effect of temperature and pressure is a prerequisite for a full understanding of the dynamic behaviour of biomolecules.

NOTES

- Gamma radiation of cellular water rapidly generates the Reactive Oxygen Species (ROS) Hydroxyl Radical ($^{\bullet}OH$) and ionised water (H_2O^+), as well as the less investigated reductants Hydrogen Radical ($H^{\bullet}$) and Hydrated Electrons (e_{aq}^-). Within one ps (10^{-12} s), Superoxide ($O_2^{\bullet-}$) and Hydrogen Peroxide (H_2O_2) are formed as secondary ROS products of Infrared (IR). Subsequent chemical cascades affect the intracellular stoichiometry of these reactive species and generate additional cell-damaging molecules.

Check Your Progress

1. Explain about the natural radiation.
2. What do you understand by cosmic rays?
3. What is light?
4. Define the reflection of light.
5. Analyse the absorption of light.
6. Comprehend the energy state of an atom.
7. What is spin of electron?
8. Distinguish between the ground state and excited state of atom.
9. Define the biomolecules.

3.9 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Natural radiation is a measure of the level of ionising radiation present in the environment at a particular location which is not due to deliberate introduction of radiation sources. Natural radiation originates from a variety of sources, both natural and artificial. These include both cosmic radiation and environmental radioactivity from naturally occurring radioactive materials (such as radon and radium), as well as man-made medical X-rays, fallout from nuclear weapons testing and nuclear accidents.
2. Cosmic radiation is an ionising radiation produced when primary photons and α-particles from outside the solar system interact with components of the earth's atmosphere. A second source of cosmic radiation is the release of charged particles from the sun, which become significant during periods of solar flare ('Sun Storm').
3. Light is a form of energy which is given by luminous objects. The Sun, bulb, candle, etc., are luminous objects. Other objects which do not give out light are called non luminous objects. Light can pass through transparent materials. Glass, water, clear plastic and air are transparent materials. Light cannot pass through translucent materials.

NOTES

4. Reflection is the change in direction of a wave front at an interface between two different media so that the wave front returns into the medium from which it originated.
5. Light absorption is a process by which light is absorbed and converted into energy. An example of this process is photosynthesis in plants.
6. A quantum mechanical system or particle that is bound it means that, confined spatially can only take on certain discrete values of energy, called energy state of atom.
7. Spin is an intrinsic form of angular momentum carried by elementary particles, composite particles (hadrons), and atomic nuclei. Spin is one of two types of angular momentum in quantum mechanics, the other being orbital angular momentum. The orbital angular momentum operator is the quantum-mechanical counterpart to the classical angular momentum of orbital revolution and appears when there is periodic structure to its wave function as the angle varies.
8. The ground state of a quantum-mechanical system is its lowest-energy state; the energy of the ground state is known as the zero-point energy of the system. While molecule absorbed energy it will get exited position. An excited state is any state with energy greater than the ground state.
9. A biomolecule or biological molecule is a loosely used term for molecules present in organisms that are essential to one or more typically biological processes, such as cell division, morphogenesis, or development. Biomolecules include large macromolecules (or polyanions), such as proteins, carbohydrates, lipids, and nucleic acids, as well as small molecules, such as primary metabolites, secondary metabolites and natural products.

3.10 SUMMARY

- Natural radiation is a measure of the level of ionising radiation present in the environment at a particular location which is not due to deliberate introduction of radiation sources. Background radiation originates from a variety of sources, both natural and artificial.
- Cosmic radiation is an ionising radiation produced when primary photons and α-particles from outside the solar system interact with components of the earth's atmosphere. A second source of cosmic radiation is the release of charged particles from the sun, which become significant during periods of solar flare ('Sun Storm').
- Terrestrial radiation refers to sources of radiation that are in the soil, water, and vegetation. The major isotopes of concern for terrestrial radiation are Potassium, Uranium and the decay products of Uranium, such as Thorium, Radium, and Radon.

NOTES

- Light is a form of energy which is given by luminous objects. The Sun, bulb, candle, etc., are luminous objects. Other objects which do not give out light are called non-luminous objects. Light can pass through transparent materials. Glass, water, clear plastic and air are transparent materials.
- Reflection is the change in direction of a wave front at an interface between two different media so that the wave front returns into the medium from which it originated.
- In physics, interference is a phenomenon in which two waves superpose to form a resultant wave of greater, lower, or the same amplitude. Constructive and destructive interference result from the interaction of waves that are correlated or coherent with each other, either because they come from the same source or because they have the same or nearly the same frequency.
- The absorption of light is therefore directly proportional to the frequency. If they are complementary, light is absorbed. If they are not complementary, then the light passes through the object or gets reflected.
- Absorption depends on the state of an object's electron. All electrons vibrate at a specific frequency, which is known as their 'Natural' frequency. When light interacts with an atom of the same frequency, the electrons of the atom become excited and start vibrating.
- By Relying on this method, physicists are able to determine and identify the properties and material composition of an object by observing which frequencies of light it absorbs. While some materials are opaque to some wavelengths of light, they are transparent to others.
- A quantum mechanical system or particle that is bound it means that, confined spatially can only take on certain discrete values of energy, called energy state of atom.
- In chemistry and atomic physics, an electron shell, or principal energy level, may be thought of as the orbit of one or more electrons around an atom's nucleus.
- If the potential energy is set to zero at infinite distance from the atomic nucleus or molecule, the usual convention, then bound electron states have negative potential energy.
- Spin is an intrinsic form of angular momentum carried by elementary particles, composite particles (hadrons), and atomic nuclei. Spin is one of two types of angular momentum in quantum mechanics, the other being orbital angular momentum.
- The System International (SI) unit of spin is the same as classical angular momentum (i.e., $N{\cdot}m{\cdot}s$ or $kg{\cdot}m^2{\cdot}s^{-1}$). In practice, spin is given as a dimensionless spin quantum number by dividing the spin angular momentum by the reduced Planck constant ‘, which has the same dimensions as angular momentum, although this is not the full computation of this value.

NOTES

- In atomic physics, the spin quantum number is a quantum number (designated m_s) which describes the intrinsic angular momentum (or spin angular momentum, or simply spin) of an electron or other particle.
- The ground state of a quantum-mechanical system is its lowest-energy state; the energy of the ground state is known as the zero-point energy of the system. While molecule absorbed energy it will get exited position.
- In quantum field theory, the ground state is usually called the vacuum state or the vacuum. If more than one ground state exists, they are said to be degenerate. Many systems have degenerate ground states.
- The ground state of the hydrogen atom corresponds to having the atom's single electron in the lowest possible orbital (that is, the spherically symmetric '1s' wave function, which, so far, has demonstrated to have the lowest possible quantum numbers).
- A biomolecule or biological molecule is a loosely used term for molecules present in organisms that are essential to one or more typically biological processes, such as cell division, morphogenesis, or development. Biomolecules include large macromolecules (or polyanions), such as proteins, carbohydrates, lipids, and nucleic acids, as well as small molecules, such as primary metabolites, secondary metabolites and natural products.
- Biology and its subfields of biochemistry and molecular biology study biomolecules and their reactions. Most biomolecules are organic compounds, and just four elements oxygen, carbon, hydrogen, and nitrogen make up 96% of the human body's mass. But many other elements, such as the various biometals, are also present in small amounts.
- Biomolecules are known to play a crucial role in the growth and microstructure formation of some biological nanocomposites that exhibit superior mechanical properties. Thus, biomolecules are considered promising compounds to manipulate the structure of bio-inspired materials.

3.11 KEY WORDS

- **Natural radiation:** Natural radiation is a measure of the level of ionising radiation present in the environment at a particular location which is not due to deliberate introduction of radiation sources. Natural radiation originates from a variety of sources, both natural and artificial.
- **Light:** Light is a form of energy which is given by luminous objects. The Sun, bulb, candle, etc., are luminous objects. Other objects which do not give out light are called non-luminous objects. Light can pass through transparent materials.
- **Diffraction of light:** Diffraction refers to various phenomena that occur when a wave encounters an obstacle or opening. It is defined as the bending

of waves around the corners of an obstacle or through an aperture into the region of geometrical shadow of the obstacle/aperture.

- **Polarisation of light:** Polarisation is a property applying to transverse waves that specifies the geometrical orientation of the oscillations.
- **Scattering of light:** Scattering is a term used in physics to describe a wide range of physical processes where moving particles or radiation of some form, such as light or sound, is forced to deviate from a straight trajectory by localized non-uniformities (including particles and radiation) in the medium through which they pass.
- **Ground state:** The ground state of a quantum-mechanical system is its lowest-energy state; the energy of the ground state is known as the zero-point energy of the system.

NOTES

3.12 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term natural radiation.
2. Explain the cosmic radiation.
3. What is light?
4. Give the properties of light.
5. What do you understand by absorption of light?
6. Elaborate on energy state of atom.
7. Explain about the spin properties of electron.
8. Illustrate the ground state and excited state of atom giving figure.
9. What is biomolecules?
10. Elaborate on the effects of biomolecules.

Long-Answer Questions

1. Briefly explain about the natural radiation.
2. Discuss in detail the properties of light.
3. Describe the absorption of light.
4. Analyse the energy state of atom.
5. Explain in detail about the spin property of electron.
6. What is energy state of atom? Briefly discuss the ground state and excited state of atom giving appropriate examples.
7. Describe the biomolecule and their effects.

NOTES

3.13 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 4 SPECTROSCOPY

NOTES

Structure

4.0 INTRODUCTION

Spectroscopy is the study of the interaction between matter and electromagnetic radiation as a function of the wavelength or frequency of the radiation. Spectroscopy is used for the measurement of radiation intensity as a function of wavelength and are often used to describe experimental spectroscopic methods. Spectral measurement devices are referred to as spectrometers, spectrophotometers, spectrographs or spectral analysers.

Basically all the recognised biologic changes that are produced by means of ionizing radiation are referred as 'Delayed Effects', since they develop only after a significant latent period. Delayed effects of radiation can be explained with reference to the health hazards and injuries caused due to radiation.

The units of measurement for radioactivity are the Becquerel (Bq, international unit) and the Curie (Ci, U.S. unit). Radioactivity refers to the amount of ionising radiation released by a material. Whether it emits alpha or beta particles, gamma rays, x-rays, or neutrons, a quantity of radioactive material is expressed in terms of its radioactivity (or simply its activity). This represents how many atoms in the material decay in a given time period.

A Geiger counter is an instrument used for detecting and measuring ionising radiation. Also known as a Geiger–Müller counter (or Geiger–Muller counter), it is widely used in applications, such as radiation dosimetry, radiological protection, experimental physics, and the nuclear industry.

NOTES

Radioactive isotopes and radioactively labelled molecules are used as tracers to identify abnormal bodily processes. This is possible because some elements tend to concentrate (in compound form) in certain parts of the body – iodine in the thyroid, phosphorus in the bones and potassium in the muscles.

An autoradiograph is an image on an x-ray film or nuclear emulsion produced by the pattern of decay emissions (e.g., beta particles or gamma rays) from a distribution of a radioactive substance.

In this unit, you will study about the spectroscopy-principle and applications, delayed effect of radiation, measurement of radioactivity, Geiger Muller counter, isotopes as a tracers, autoradiography.

4.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand what spectroscopy is
- Explain the delayed effect of radiation
- Discuss about the measurement of radioactivity
- Analyse the Geiger Muller counter
- Elaborate on the isotopes as a tracers
- Interpret the autoradiography

4.2 SPECTROSCOPY: PRINCIPLES AND APPLICATIONS

Spectroscopy is the study of the interaction between matter and electromagnetic radiation as a function of the wavelength or frequency of the radiation. In simpler terms, spectroscopy is the precise study of colour as generalised from visible light to all bands of the electromagnetic spectrum; indeed, historically, spectroscopy originated as the study of the wavelength dependence of the absorption by gas phase matter of visible light dispersed by a prism. Matter waves and acoustic waves can also be considered forms of radiative energy, and recently gravitational waves have been associated with a spectral signature in the context of the Laser Interferometer Gravitational-Wave Observatory (LIGO).

Spectroscopy, primarily in the electromagnetic spectrum, is a fundamental exploratory tool in the fields of physics, chemistry, and astronomy, allowing the composition, physical structure and electronic structure of matter to be investigated at the atomic, molecular and macro scale, and over astronomical distances. Important applications arise from biomedical spectroscopy in the areas of tissue analysis and medical imaging.

Spectroscopy and spectrographic technique are terms used to refer to the measurement of radiation intensity as a function of wavelength and are often used to describe experimental spectroscopic methods. Spectral measurement devices are referred to as spectrometers, spectrophotometers, spectrographs or spectral analysers.

Daily observations of colour can be related to spectroscopy. Neon lighting is a direct application of atomic spectroscopy. Neon and other noble gases have characteristic emission frequencies (colours). Neon lamps use collision of electrons with the gas to excite these emissions. Inks, dyes and paints include chemical compounds selected for their spectral characteristics in order to generate specific colours and hues. A commonly encountered molecular spectrum is that of nitrogen dioxide. Gaseous nitrogen dioxide has a characteristic red absorption feature, and this gives air polluted with nitrogen dioxide a reddish-brown colour. Rayleigh scattering is a spectroscopic scattering phenomenon that accounts for the colour of the sky.

Spectroscopic studies were central to the development of quantum mechanics and included Max Planck's explanation of blackbody radiation, Albert Einstein's explanation of the photoelectric effect and Niels Bohr's explanation of atomic structure and spectra. Spectroscopy is used in physical and analytical chemistry because atoms and molecules have unique spectra. As a result, these spectra can be used to detect, identify and quantify information about the atoms and molecules. Spectroscopy is also used in astronomy and remote sensing on Earth. Most research telescopes have spectrographs. The measured spectra are used to determine the chemical composition and physical properties of astronomical objects (such as, their temperature and velocity).

Principles of Spectroscopy: Spectroscopy deals with the production, measurement, and interpretation of spectra arising from the interaction of electromagnetic radiation with matter. There are many different spectroscopic methods available for solving a wide range of analytical problems. The methods differ with respect to the species to be analysed (such as, molecular or atomic spectroscopy), the type of radiation-matter interaction to be monitored (such as, absorption, emission, or diffraction), and the region of the electromagnetic spectrum used in the analysis. Spectroscopic methods are very informative and widely used for both quantitative and qualitative analyses. Spectroscopic methods based on the absorption or emission of radiation in the Ultraviolet (UV), Visible (Vis), Infrared (IR), and radio (Nuclear Magnetic Resonance, NMR) frequency ranges are most commonly encountered in traditional food analysis laboratories. Each of these methods is distinct in that it monitors different types of molecular or atomic transitions.

NOTES

Applications of Spectroscopy

There are several applications to spectroscopy in the field of medicine, physics, chemistry, and astronomy. Taking advantage of the properties of absorbance, spectroscopy can be used to identify certain states of nature. Such examples include:

- Cure monitoring of composites using optical fibers.
- Estimate weathered wood exposure times using near infrared spectroscopy.
- Measurement of different compounds in food samples by absorption spectroscopy both in visible and infrared spectrum.
- Measurement of toxic compounds in blood samples
- Non-destructive elemental analysis by X-ray fluorescence.
- Electronic structure research with various spectroscopes.
- Radar to determine the speed and velocity of a distant object
- Finding the physical properties of a distant star or nearby exoplanet using the Relativistic Doppler effect.

4.3 DELAYED EFFECTS OF RADIATIONS

Radiation and Delayed Effect of Radiation

In physics, radiation is the emission or transmission of energy in the form of waves or particles through space or through a material medium. This includes the following:

- Electromagnetic radiation, such as radio waves, microwaves, infrared, visible light, ultraviolet, X-rays, and gamma radiation (γ).
- Particle radiation, such as alpha radiation (α), beta radiation (β), proton radiation and neutron radiation (particles of non-zero rest energy).
- Acoustic radiation, such as ultrasound, sound, and seismic waves (dependent on a physical transmission medium).
- Gravitational radiation, radiation that takes the form of gravitational waves, or ripples in the curvature of space time.

Radiation is often categorized as either ionizing or non-ionizing depending on the energy of the radiated particles. Ionizing radiation carries more than 10 eV, which is enough to ionize atoms and molecules and break chemical bonds. This is an important distinction due to the large difference in harmfulness to living organisms. A common source of ionizing radiation is radioactive materials that emit α, β, or γ radiation, consisting of helium nuclei, electrons or positrons, and photons, respectively. Other sources include X-rays from medical radiography examinations and muons, mesons, positrons, neutrons and other particles that constitute the

secondary cosmic rays that are produced after primary cosmic rays interact with Earth's atmosphere.

Gamma rays, X-rays and the higher energy range of ultraviolet light constitute the ionizing part of the electromagnetic spectrum. The word 'Ionize' refers to the breaking of one or more electrons away from an atom, an action that requires the relatively high energies that these electromagnetic waves supply. Further down the spectrum, the non-ionizing lower energies of the lower ultraviolet spectrum cannot ionize atoms, but can disrupt the inter-atomic bonds which form molecules, thereby breaking down molecules rather than atoms; a good example of this is sunburn caused by long-wavelength solar ultraviolet. The waves of longer wavelength than UV in visible light, infrared and microwave frequencies cannot break bonds but can cause vibrations in the bonds which are sensed as heat. Radio wavelengths and below generally are not regarded as harmful to biological systems.

The word radiation arises from the phenomenon of waves radiating (i.e., traveling outward in all directions) from a source. This aspect leads to a system of measurements and physical units that are applicable to all types of radiation. Because such radiation expands as it passes through space, and as its energy is conserved (in vacuum), the intensity of all types of radiation from a point source follows an inverse-square law in relation to the distance from its source. Like any ideal law, the inverse-square law approximates a measured radiation intensity to the extent that the source approximates a geometric point.

Radiation and radioactive substances are used for diagnosis, treatment, and research. X-rays, for example, pass through muscles and other soft tissue but are stopped by dense materials. This property of X-rays enables doctors to find broken bones and to locate cancers that might be growing in the body. Doctors also find certain diseases by injecting a radioactive substance and monitoring the radiation given off as the substance moves through the body. Radiation used for cancer treatment is called ionizing radiation because it forms ions in the cells of the tissues it passes through as it dislodges electrons from atoms. This can kill cells or change genes so the cells cannot grow. Other forms of radiation, such as radio waves, microwaves, and light waves are called non-ionizing. They do not have as much energy so they are not able to ionize cells.

Possible Damage to Health from Certain Types of Radiation

Ionizing radiation in certain conditions can cause damage to living organisms, causing cancer or genetic damage.

Non-ionizing radiation in certain conditions also can cause damage to living organisms, such as burns. In 2011, the International Agency for Research on Cancer (IARC) of the World Health Organization (WHO) released a statement adding radio frequency electromagnetic fields (including microwave and millimeter waves) to their list of things which are possibly carcinogenic to humans.

NOTES

Delayed Effect of Radiation

Basically all the recognised biologic changes that are produced by means of ionizing radiation are referred as 'Delayed Effects', since they develop only after a significant latent period. Delayed effects of radiation can be explained with reference to the health hazards and injuries caused due to radiation. It is expected, however, to limit the term 'Delayed Effects' to those injuries that first become apparent several months or years after exposure. Tissues in which delayed effects are destined to appear seldom show a steady progression of injury from the early to the late reaction. Characteristically, the acute response subsides and the tissues resume a fairly normal appearance or at least reach a stable state of relative ischemia. At any time thereafter the late lesions may develop.

With any exposure to radiation, there is some risk. Following are the approximate risks for the three principal effects of exposure to low levels of radiation:

Effect	Excess Cases per 10,000 Exposed Per Rad
Genetic	2 to 4
Somatic (Cancer)	4 to 20
In-Utero (Cancer)	4 to 12
In-Utero (All Effects)	20 to 200

Genetic - Risks from 1 rem of radiation exposure to the reproductive organs are approximately 50 to 1,000 times less than the spontaneous risk for various anomalies.

Somatic - For radiation induced cancers, the risk approximation is lesser compared to the normal incidence of about 1 in 4 chances of developing any type of cancer. However, not all cancers are associated with exposure to radiation. The risk of dying from radiation induced cancer is about one half the risk of getting the cancer.

In-Utero - Spontaneous risks of fetal abnormalities are about 5 to 30 times greater than the risk of exposure to 1 rem of radiation. However, the risk of childhood cancer from exposure in-utero is similar as the risk to adults exposed to radiation. By far, medical practice is the largest source of in-utero radiation exposure.

The increased use of radioisotopes has led to increased concerns over the effects of these materials on biological systems (such as, humans). All radioactive nuclides emit high-energy particles or electromagnetic waves. When this radiation encounters living cells, it can cause heating, break chemical bonds, or ionize molecules. The most serious biological damage results when these radioactive emissions fragment or ionize molecules. For example, alpha and beta particles emitted from nuclear decay reactions possess much higher energies than ordinary chemical bond energies. When these particles strike and penetrate matter, they produce ions and molecular fragments that are extremely reactive. This damage

happens to biomolecules in living organisms and can cause serious malfunctions in normal cell processes, possibly causing illness or even death.

Most delayed radiation injuries are the consequence of accidental or misguided overexposure. Radiation injuries may be divided into two key categories, according to whether they are local or general in nature. Localized lesions usually result from local irradiation but some local lesions, such as cataract, may also follow whole-body exposure. In similar method, generalized effects are seen most often when either a large area or the entire body has been exposed, though considerable local treatment at times produces systemic effects.

Late radiation ulcers usually result from localized radiation therapy. They occur most often in the skin but are also observed in the rectum and bladder and occasionally elsewhere. Even though the dose has been of such a size that late lesions will ultimately develop, the early acute reaction usually heals, leaving a poorly vascularized tissue which is often the seat of telangiectases, chronic edema, and fibrosis. Ulceration begins after an indeterminate latent period of months or years and extends progressively to produce a sharply punched-out ulcer crater with a brawny, indurated, ischemic base. Once the ulcer has developed, it heals poorly if at all, and even if it heals is prone to recur. Ischemia appears to be a major underlying factor in the pathogenesis of ulcerating lesions, although the onset is often precipitated by some secondary injury.

Chronic radiation dermatitis, resembling in many respects spontaneous senile degeneration of the skin, may persist for many years without ulceration.

4.4 MEASUREMENTS OF RADIOACTIVITY

The local presence of nuclear radiation arising from the radioactive decay of radionuclides. The unit of radioactivity from the System International of units (SI system) is the Becquerel (Bq) defined as the radioactive decay or disintegration of one radionuclide per second. Radioactivity is a measure of the total, local rate of radionuclides decaying per unit time and is dependent upon the total number of atoms, decay constants, and all decay branching pathways for each radionuclide present.

The Becquerel (Bq) is the SI derived unit of radioactivity. One Becquerel is defined as the activity of a quantity of radioactive material in which one nucleus decays per second. For applications relating to human health this is a small quantity, and SI multiples of the unit are commonly used.

The Becquerel is named after Henri Becquerel, who shared a Nobel Prize in Physics with Pierre and Marie Skłodowska Curie in 1903 for their work in discovering radioactivity.

Definition

$1 \text{ Bq} = 1 \text{ s}^{-1}$

NOTES

A special name was introduced for the reciprocal second (s^{-1}) to represent radioactivity to avoid potentially dangerous mistakes with prefixes. For example, 1 μs^{-1} would mean 106 disintegrations per second: $1{\cdot}(10^{-6}\ s)^{-1} = 10^6\ s^{-1}$, whereas 1 μBq would mean 1 disintegration per 1 million seconds. Other names considered were Hertz (Hz), a special name already in use for the reciprocal second, and Fourier (Fr). The hertz is now only used for periodic phenomena. Whereas 1 Hz is 1 cycle per second, 1 Bq is 1 aperiodic radioactivity event per second.

The Gray (Gy) and the Becquerel (Bq) were introduced in 1975. Between 1953 and 1975, absorbed dose was often measured in rads. Decay activity was measured in curies before 1946 and often in Rutherford's between 1946 and 1975.

Unit Capitalisation

As with every International System of Units (SI) unit named for a person, the first letter of its symbol is uppercase (Bq).

Like any SI unit, Bq can be prefixed; commonly used multiples are kBq (kilo Becquerel, 10^3 Bq), MBq (Mega Becquerel, 10^6 Bq, equivalent to 1 Rutherford), GBq (Giga Becquerel, 10^9 Bq), TBq (Tera Becquerel, 10^{12} Bq), and PBq (Peta Becquerel, 10^{15} Bq). Large prefixes are common for practical uses of the unit.

Calculation of Radioactivity

For a given mass m (in grams) of an isotope with atomic mass m_a (in g/mol) and a half-life of $t_{1/2}$ (in s), the radioactivity can be calculated using:

$$A_{Bq} = \frac{m}{m_a} N_A \frac{1n2}{t_{\frac{1}{2}}}$$

With $N_A = 6.02214076{\times}10^{23}\ mol^{-1}$, the Avogadro constant.

Since m/m_a is the number of moles (n), the amount of radioactivity A can be calculated by:

$$A_{Bq} = nN_A \frac{1n2}{t_{\frac{1}{2}}}$$

For instance, on average each gram of potassium contains 0.000117 gram of ^{40}K (all other naturally occurring isotopes are stable) that has a $t_{1/2}$ of $1.277{\times}10^9$ years = $4.030{\times}10^{16}$ s, and has an atomic mass of 39.964 g/mol, so the amount of radioactivity associated with a gram of potassium is 30 Bq.

Examples

For practical applications, 1 Bq is a small unit. For example, the roughly 0.0169 g of potassium-40 present in a typical human body produces approximately 4,400 disintegrations per second or 4.4 kBq of activity.

The global inventory of carbon-14 is estimated to be 8.5×10^{18} Bq (8.5 E Bq, 8.5 Exa Becquerel). The nuclear explosion in Hiroshima (an explosion of 16 kt or 67 TJ) is estimated to have injected 8×10^{24} Bq (8 Y Bq, 8 Yotta Becquerel) of radioactive fission products into the atmosphere.

These examples are useful for comparing the amount activity of these radioactive materials but should not be confused with the amount of exposure to ionising radiation that these materials represent. The level of exposure and thus the absorbed dose received are what should be considered when assessing the effects of ionising radiation on humans.

Relation to the Curie

The Becquerel succeeded the Curie (Ci), an older, non-SI unit of radioactivity based on the activity of 1 gram of radium-226. The curie is defined as 3.7×10^{10} s^{-1}, or 37 GBq.

Conversion Factors:

1 Ci = 3.7×10^{10} Bq = 37 GBq

1 μCi = 37,000 Bq = 37 kBq

1 Bq = 2.7×10^{-11} Ci = 2.7×10^{-5} μCi

1 MBq = 0.027 mCi

Relation to other Radiation-Related Quantities

The following table shows radiation quantities in SI and non-SI units. W_R (formerly 'Q' factor) is a factor that scales the biological effect for different types of radiation, relative to x-rays. (e.g. 1 for beta radiation, 20 for alpha radiation, and a complicated function of energy for neutrons) In general conversion between rates of emission, the density of radiation, the fraction absorbed, and the biological effects, requires knowledge of the geometry between source and target, the energy and the type of the radiation emitted, among other factors.

Table 4.1 *Ionising Radiation Related Quantities*

Quantity	Unit	Symbol	Derivation	Year	SI equivalence
Activity (*A*)	Becquerel	Bq	s^{-1}	1974	SI unit
	Curie	Ci	3.7×10^{10} s^{-1}	1953	3.7×10^{10} Bq
	Rutherford	Rd	10^6 s^{-1}	1946	1,000,000 Bq
Exposure (*X*)	Coulomb per kilogram	C/kg	$C{\cdot}kg^{-1}$ of air	1974	SI unit
	Röntgen	R	esu / 0.001293 g of air	1928	2.58×10^{-4} C/kg
Absorbed dose (*D*)	Gray	Gy	$J{\cdot}kg^{-1}$	1974	SI unit
	erg per gram	erg/g	$erg{\cdot}g^{-1}$	1950	1.0×10^{-4} Gy
	Rad	rad	$100\ erg{\cdot}g^{-1}$	1953	0.010 Gy
Equivalent dose (*H*)	Sievert	Sv	$J{\cdot}kg^{-1} \times W_R$	1977	SI unit
	Röntgen equivalent man	rem	$100\ erg{\cdot}g^{-1}$ x W_R	1971	0.010 Sv
Effective dose (*E*)	Sievert	Sv	$J{\cdot}kg^{-1} \times W_R \times W_T$	1977	SI unit
	Röntgen equivalent man	rem	$100\ erg{\cdot}g^{-1} \times W_R \times W_T$	1971	0.010 Sv

4.5 GEIGER MULLER COUNTER

NOTES

A Geiger counter is an instrument used for detecting and measuring ionising radiation. Also known as a Geiger–Müller counter (or Geiger–Muller counter), it is widely used in applications, such as radiation dosimetry, radiological protection, experimental physics, and the nuclear industry.

It detects ionizing radiation, such as alpha particles, beta particles, and gamma rays using the ionisation effect produced in a Geiger–Müller tube, which gives its name to the instrument. In wide and prominent use as a hand-held radiation survey instrument, it is perhaps one of the world's best-known radiation detection instruments.

The original detection principle was realised in 1908 at the University of Manchester, but it was not until the development of the Geiger–Müller tube in 1928 that the Geiger counter could be produced as a practical instrument. Since then, it has been very popular due to its robust sensing element and relatively low cost. However, there are limitations in measuring high radiation rates and the energy of incident radiation.

Principle of Operation

A Geiger counter consists of a Geiger–Müller tube (the sensing element which detects the radiation) and the processing electronics, which displays the result.

The Geiger–Müller tube is filled with an inert gas, such as helium, neon, or argon at low pressure, to which a high voltage is applied. The tube briefly conducts electrical charge when a particle or photon of incident radiation makes the gas conductive by ionisation. The ionisation is considerably amplified within the tube by the Townsend discharge effect to produce an easily measured detection pulse, which is fed to the processing and display electronics. This large pulse from the tube makes the Geiger counter relatively cheap to manufacture, as the subsequent electronics are greatly simplified. The electronics also generate the high voltage, typically 400–900 volts that has to be applied to the Geiger–Müller tube to enable its operation. To stop the discharge in the Geiger–Müller tube a little halogen gas or organic material (alcohol) is added to the gas mixture.

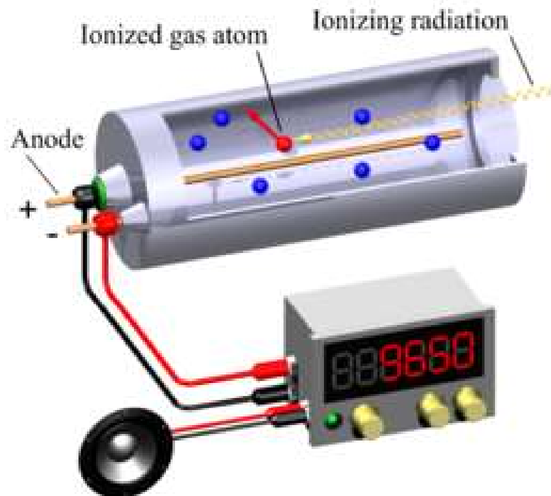


Fig. 4.1 *Diagram of a Geiger Counter*

Limitations

There are two main limitations of the Geiger counter. Because the output pulse from a Geiger–Müller tube is always of the same magnitude (regardless of the energy of the incident radiation), the tube cannot differentiate between radiation types. Secondly, the tube cannot measure high radiation rates, because each ionisation event is followed by a 'Dead Time', an insensitive period during which any further incident radiation does not result in a count. Typically, the dead time will reduce indicated count rates above about 104 to 105 counts per second, depending on the characteristic of the tube being used. While some counters have circuitry which can compensate for this, for accurate measurements ion chamber instruments are preferred for high radiation rates.

Types and Applications

The intended detection application of a Geiger counter dictates the tube design used. Consequently, there are a great many designs, but they can be generally categorized as 'End-Window', windowless 'Thin-Walled', 'Thick-Walled', and sometimes hybrids of these types.

Particle Detection: The first historical uses of the Geiger principle were for the detection of alpha and beta particles, and the instrument is still used for this purpose today. For alpha particles and low energy beta particles, the end-window type of a Geiger–Müller tube has to be used as these particles have a limited range and are easily stopped by a solid material. Therefore, the tube requires a window which is thin enough to allow as many as possible of these particles through to the fill gas. The window is usually made of mica with a density of about 1.5 - 2.0 mg/cm^2.

Alpha particles have the shortest range, and to detect these the window should ideally be within 10 mm of the radiation source due to alpha particle attenuation. However, the Geiger–Müller tube produces a pulse output which is the same magnitude for all detected radiation, so a Geiger counter with an end window tube cannot distinguish between alpha and beta particles.A skilled operator

NOTES

can use varying distance from a radiation source to differentiate between alpha and high energy beta particles.

The Pancake' Geiger–Müller tube is a variant of the end-window probe, but designed with a larger detection area to make checking quicker. However, the pressure of the atmosphere against the low pressure of the fill gas limits the window size due to the limited strength of the window membrane.

Some beta particles can also be detected by a thin-walled 'Windowless' Geiger–Müller tube, which has no end-window, but allows high energy beta particles to pass through the tube walls. Although the tube walls have a greater stopping power than a thin end-window, they still allow these more energetic particles to reach the fill gas.

Gamma and X-ray Detection: Geiger counters are widely used to detect gamma radiation and X-rays collectively known as photons, and for this the windowless tube is used. However, detection efficiency is low compared to alpha and beta particles. The article on the Geiger–Müller tube carries a more detailed account of the techniques used to detect photon radiation. For high energy photons the tube relies on the interaction of the radiation with the tube wall, usually a high Z material, such as chrome steel of 1–2 mm thickness to produce electrons within the tube wall. These enter and ionize the fill gas.

This is necessary as the low-pressure gas in the tube has little interaction with higher energy photons. However, as photon energies decrease to low levels there is greater gas interaction and the direct gas interaction increases. At very low energies (less than 25 KeV) direct gas ionisation dominates and a steel tube attenuates the incident photons. Consequently, at these energies, a typical tube design is a long tube with a thin wall which has a larger gas volume to give an increased chance direct interaction of a particle with the fill gas.

Above these low energy levels, there is a considerable variance in response to different photon energies of the same intensity, and a steel-walled tube employs what is known as 'Energy Compensation' in the form of filter rings around the naked tube which attempts to compensate for these variations over a large energy range. A chrome steel G-M tube is about 1% efficient over a wide range of energies.

Neutron Detection: A variation of the Geiger tube is used to measure neutrons, where the gas used is Boron Trifluoride or Helium-3 and a plastic moderator is used to slow the neutrons. This creates an alpha particle inside the detector and thus neutrons can be counted.

4.6 ISOTOPES AS TRACERS

A radioactive tracer, radiotracer, or radioactive label, is a chemical compound in which one or more atoms have been replaced by a radionuclide so by virtue of its radioactive decay it can be used to explore the mechanism of chemical reactions

by tracing the path that the radioisotope follows from reactants to products. Radiolabelling or radio tracing is thus the radioactive form of isotopic labelling.

Radioisotopes of hydrogen, carbon, phosphorus, sulphur, and iodine have been used extensively to trace the path of biochemical reactions. A radioactive tracer can also be used to track the distribution of a substance within a natural system such as a cell or tissue, or as a flow tracer to track fluid flow. Radioactive tracers are also used to determine the location of fractures created by hydraulic fracturing in natural gas production. Radioactive tracers form the basis of a variety of imaging systems, such as, Polyethylene Terephthalate (PET) scans, Single-Photon Emission Computerised Tomography (SPECT) scans and technetium scans. Radiocarbon dating uses the naturally occurring carbon-14 isotope as an isotopic label.

Methodology: Isotopes of a chemical element differ only in the mass number. For example, the isotopes of hydrogen can be written as ^{1}H, ^{2}H and ^{3}H, with the mass number superscripted to the left. When the atomic nucleus of an isotope is unstable, compounds containing this isotope are radioactive. Tritium is an example of a radioactive isotope.

The principle behind the use of radioactive tracers is that an atom in a chemical compound is replaced by another atom, of the same chemical element. The substituting atom, however, is a radioactive isotope. This process is often called radioactive labelling. The power of the technique is due to the fact that radioactive decay is much more energetic than chemical reactions. Therefore, the radioactive isotope can be present in low concentration and its presence detected by sensitive radiation detectors, such as Geiger counters and scintillation counters. George de Hevesy won the 1943 Nobel Prize for Chemistry for his work on the use of isotopes as tracers in the study of chemical processes.

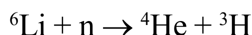
There are two main ways in which radioactive tracers are used:

1. When a labelled chemical compound undergoes chemical reactions one or more of the products will contain the radioactive label. Analysis of what happens to the radioactive isotope provides detailed information on the mechanism of the chemical reaction.
2. A radioactive compound is introduced into a living organism and the radio-isotope provides a means to construct an image showing the way in which that compound and its reaction products are distributed around the organism.

Tracer Isotopes

Hydrogen

Tritium is produced by neutron irradiation of ^{6}Li



$^{6}Li + n \rightarrow {}^{4}He + {}^{3}H$

Tritium has a half-life 4,500±8 days (approximately 12.32 years), and it decays by beta decay. The electrons produced have an average energy of 5.7

keV. Because the emitted electrons have relatively low energy, the detection efficiency by scintillation counting is rather low. However, hydrogen atoms are present in all organic compounds, so tritium is frequently used as a tracer in biochemical studies.

NOTES

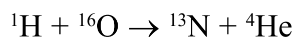
Carbon

^{11}C decays by positron emission with a half-life of ca. 20 min. ^{11}C is one of the isotopes often used in positron emission tomography.

^{14}C decays by beta decay, with a half-life of 5730 years. It is continuously produced in the upper atmosphere of the earth, so it occurs at a trace level in the environment. However, it is not practical to use naturally-occurring ^{14}C for tracer studies. Instead it is made by neutron irradiation of the isotope ^{13}C which occurs naturally in carbon at about the 1.1% level. ^{14}C has been used extensively to trace the progress of organic molecules through metabolic pathways.

Nitrogen

^{13}N decays by positron emission with a half-life of 9.97 min. It is produced by the nuclear reaction



$^{1}H + {}^{16}O \rightarrow {}^{13}N + {}^{4}He$

^{13}N is used in Positron Emission Tomography (PET scan).

Oxygen

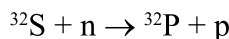
^{15}O decays by positron emission with a half-life of 122 sec. It is used in positron emission tomography.

Fluorine

^{18}F decays predominately by β-emission, with a half-life of 109.8 min. It is made by proton bombardment of ^{18}O in a cyclotron or linear particle accelerator. It is an important isotope in the radiopharmaceutical industry. For example, it is used to make labelled Fluoro Deoxy Glucose (FDG) for application in PET scans.

Phosphorus

^{32}P is made by neutron bombardment of ^{32}S



$^{32}S + n \rightarrow {}^{32}P + p$

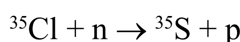
It decays by beta decay with a half-life of 14.29 days. It is commonly used to study protein phosphorylation by kinases in biochemistry.

^{33}P is made in relatively low yield by neutron bombardment of ^{31}P. It is also a beta-emitter, with a half-life of 25.4 days. Though more expensive than ^{32}P, the emitted electrons are less energetic, permitting better resolution in, for example Deoxyribonucleic Acid (DNA) sequencing.

Both isotopes are useful for labelling nucleotides and other species that contain a phosphate group.

Sulphur

^{35}S is made by neutron bombardment of ^{35}Cl



$^{35}Cl + n \rightarrow {}^{35}S + p$

NOTES

It decays by beta-decay with a half-life of 87.51 days. It is used to label the sulphur-containing amino-acids methionine and cysteine. When a sulphur atom replaces an oxygen atom in a phosphate group on a nucleotide a thiophosphate is produced, so ^{35}S can also be used to trace a phosphate group.

Iodine

^{123}I is produced by proton irradiation of ^{124}Xe. The caesium isotope produced is unstable and decays to ^{123}I. The isotope is usually supplied as the iodide and hypo iodate in dilute sodium hydroxide solution, at high isotopic purity. ^{123}I has also been produced at Oak Ridge National Laboratories by proton bombardment of ^{123}Te.

^{123}I decays by electron capture with a half-life of 13.22 hours. The emitted 159 keV gamma ray is used in Single-Photon Emission Computed Tomography (SPECT). A 127 keV gamma ray is also emitted.

^{125}I is frequently used in radioimmunoassay because of its relatively long half-life (59 days) and ability to be detected with high sensitivity by gamma counters.

^{129}I is present in the environment as a result of the testing of nuclear weapons in the atmosphere. It was also produced in the Chernobyl and Fukushima disasters. ^{129}I decays with a half-life of 15.7 million years, with low-energy beta and gamma emissions. It is not used as a tracer, though its presence in living organisms, including human beings, can be characterised by measurement of the gamma rays.

4.7 AUTORADIOGRAPHY

An autoradiograph is an image on an x-ray film or nuclear emulsion produced by the pattern of decay emissions (e.g., beta particles or gamma rays) from a distribution of a radioactive substance. Alternatively, the autoradiograph is also available as a digital image (digital autoradiography), due to the recent development of scintillation gas detectors or rare earth phosphorimaging systems. The film or emulsion is opposed to the labelled tissue section to obtain the autoradiograph (also called an autoradiogram). The auto- prefix indicates that the radioactive substance is within the sample, as distinguished from the case of historadiography or microradiography, in which the sample is marked using an external source. Some autoradiographs can be examined microscopically for localization of silver grains (such as on the interiors or exteriors of cells or organelles) in which the process is termed micro-autoradiography. For example, micro-autoradiography was used to examine whether atrazine was being metabolised by the hornwort plant or by epiphytic microorganisms in the biofilm layer surrounding the plant.

NOTES

Applications

In biology, this technique may be used to determine the tissue (or cell) localization of a radioactive substance, either introduced into a metabolic pathway, bound to a receptor or enzyme, or hybridised to a nucleic acid. Applications for autoradiography are broad, ranging from biomedical to environmental sciences to industry.

Receptor Autoradiography

The use of radiolabelled ligands to determine the tissue distributions of receptors is termed either in vivo or in vitro receptor autoradiography if the ligand is administered into the circulation (with subsequent tissue removal and sectioning) or applied to the tissue sections, respectively. Once the receptor density is known, in vitro autoradiography can also be used to determine the anatomical distribution and affinity of a radiolabelled drug towards the receptor. For in vitro autoradiography, radio ligand was directly applying on frozen tissue sections without administration to the subject. Thus it cannot follow the distribution, metabolism and degradation situation completely in the living body. But because target in the cryosections is widely exposed and can direct contact with radio ligand, in vitro autoradiography is still a quick and easy method to screen drug candidates, PET and SPECT ligands. The ligands are generally labeled with ^{3}H (tritium), ^{18}F (fluroine), ^{11}C (carbon) or ^{125}I (radioiodine). Compare to in vitro, ex vivo autoradiography were performed after administration of radio ligand in the body, which can decrease the artifacts and are closer to the inner environment.

The distribution of Ribonucleic acid (RNA) transcripts in tissue sections by the use of radiolabelled, complementary oligonucleotides or ribonucleic acids (riboprobes) is called in situ hybridization histochemistry. Radioactive precursors of DNA and RNA, [^{3}H]-Thymidine and [^{3}H]-Uridine respectively, may be introduced to living cells to determine the timing of several phases of the cell cycle. RNA or DNA viral sequences can also be located in this fashion. These probes are usually labelled with ^{32}P, ^{33}P, or ^{35}S. In the realm of behavioural endocrinology, autoradiography can be used to determine hormonal uptake and indicate receptor location; an animal can be injected with a radiolabelled hormone, or the study can be conducted in vitro.

Check Your Progress

1. Define the term spectroscopy.
2. Give the principle of spectroscopy.
3. Elaborate on the delayed effect of radiation.
4. What is Becquerel (Bq)?
5. Explain the Geiger Muller counter.
6. What do understand by isotopes as tracers?
7. Explain the autoradiography.

4.8 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. Spectroscopy is the study of the interaction between matter and electromagnetic radiation as a function of the wavelength or frequency of the radiation. In simpler terms, spectroscopy is the precise study of colour as generalised from visible light to all bands of the electromagnetic spectrum; indeed, historically, spectroscopy originated as the study of the wavelength dependence of the absorption by gas phase matter of visible light dispersed by a prism.
2. Spectroscopy deals with the production, measurement, and interpretation of spectra arising from the interaction of electromagnetic radiation with matter. There are many different spectroscopic methods available for solving a wide range of analytical problems. The methods differ with respect to the species to be analysed (such as molecular or atomic spectroscopy), the type of radiation-matter interaction to be monitored (such as, absorption, emission, or diffraction), and the region of the electromagnetic spectrum used in the analysis.
3 Ionizing radiation are referred as 'Delayed Effects', since they develop only after a significant latent period. Delayed effects of radiation can be explained with reference to the health hazards and injuries caused due to radiation. It is expected, however, to limit the term 'Delayed Effects' to those injuries that first become apparent several months or years after exposure. Tissues in which delayed effects are destined to appear seldom show a steady progression of injury from the early to the late reaction.
4. The Becquerel (Bq) is the SI derived unit of radioactivity. One Becquerel is defined as the activity of a quantity of radioactive material in which one nucleus decays per second.
5. A Geiger counter is an instrument used for detecting and measuring ionising radiation. Also known as a Geiger–Müller counter (or Geiger–Muller counter), it is widely used in applications, such as radiation dosimetry, radiological protection, experimental physics, and the nuclear industry. It detects ionizing radiation, such as alpha particles, beta particles, and gamma rays using the ionisation effect produced in a Geiger–Müller tube, which gives its name to the instrument.
6. A radioactive tracer, radiotracer, or radioactive label, is a chemical compound in which one or more atoms have been replaced by a radionuclide so by virtue of its radioactive decay it can be used to explore the mechanism of chemical reactions by tracing the path that the radioisotope follows from reactants to products. Radiolabelling or radio tracing is thus the radioactive form of isotopic labelling.

NOTES

7. An autoradiograph is an image on an x-ray film or nuclear emulsion produced by the pattern of decay emissions (e.g., beta particles or gamma rays) from a distribution of a radioactive substance. Alternatively, the autoradiograph is also available as a digital image (digital autoradiography), due to the recent development of scintillation gas detectors or rare earth phosphorimaging systems.

4.9 SUMMARY

- Matter waves and acoustic waves can also be considered forms of radiative energy, and recently gravitational waves have been associated with a spectral signature in the context of the Laser Interferometer Gravitational-Wave Observatory (LIGO).
- Spectroscopy, primarily in the electromagnetic spectrum, is a fundamental exploratory tool in the fields of physics, chemistry, and astronomy, allowing the composition, physical structure and electronic structure of matter to be investigated at the atomic, molecular and macro scale, and over astronomical distances. Important applications arise from biomedical spectroscopy in the areas of tissue analysis and medical imaging.
- Spectroscopy and spectrographic technique are terms used to refer to the measurement of radiation intensity as a function of wavelength and are often used to describe experimental spectroscopic methods. Spectral measurement devices are referred to as spectrometers, spectrophotometers, spectrographs or spectral analysers.
- Spectroscopic studies were central to the development of quantum mechanics and included Max Planck's explanation of blackbody radiation, Albert Einstein's explanation of the photoelectric effect and Niels Bohr's explanation of atomic structure and spectra. Spectroscopy is used in physical and analytical chemistry because atoms and molecules have unique spectra.
- The local presence of nuclear radiation arising from the radioactive decay of radionuclides. The unit of radioactivity from the System International of units (SI system) is the Becquerel (Bq) defined as the radioactive decay or disintegration of one radionuclide per second.
- The Becquerel (Bq) is the SI derived unit of radioactivity. One Becquerel is defined as the activity of a quantity of radioactive material in which one nucleus decays per second. For applications relating to human health this is a small quantity, and SI multiples of the unit are commonly used.
- The Becquerel is named after Henri Becquerel, who shared a Nobel Prize in Physics with Pierre and Marie Skłodowska Curie in 1903 for their work in discovering radioactivity.
- A Geiger counter is an instrument used for detecting and measuring ionising radiation. Also known as a Geiger–Müller counter (or Geiger–Muller

counter), it is widely used in applications, such as radiation dosimetry, radiological protection, experimental physics, and the nuclear industry.

NOTES

- Geiger counter detects ionising radiation, such as alpha particles, beta particles, and gamma rays using the ionisation effect produced in a Geiger–Müller tube, which gives its name to the instrument. In wide and prominent use as a hand-held radiation survey instrument, it is perhaps one of the world's best-known radiation detection instruments.
- A Geiger counter consists of a Geiger–Müller tube (the sensing element which detects the radiation) and the processing electronics, which displays the result. The Geiger–Müller tube is filled with an inert gas, such as helium, neon, or argon at low pressure, to which a high voltage is applied. The tube briefly conducts electrical charge when a particle or photon of incident radiation makes the gas conductive by ionisation.
- The ionisation is considerably amplified within the tube by the Townsend discharge effect to produce an easily measured detection pulse, which is fed to the processing and display electronics. This large pulse from the tube makes the Geiger counter relatively cheap to manufacture, as the subsequent electronics are greatly simplified.
- A radioactive tracer, radiotracer, or radioactive label, is a chemical compound in which one or more atoms have been replaced by a radionuclide so by virtue of its radioactive decay it can be used to explore the mechanism of chemical reactions by tracing the path that the radioisotope follows from reactants to products. Radiolabelling or radio tracing is thus the radioactive form of isotopic labelling.
- Radioisotopes of hydrogen, carbon, phosphorus, sulphur, and iodine have been used extensively to trace the path of biochemical reactions. A radioactive tracer can also be used to track the distribution of a substance within a natural system such as a cell or tissue, or as a flow tracer to track fluid flow. Radioactive tracers are also used to determine the location of fractures created by hydraulic fracturing in natural gas production.
- Radioactive tracers form the basis of a variety of imaging systems, such as, Polyethylene Terephthalate (PET) scans, Single-Photon Emission Computerised Tomography (SPECT) scans and technetium scans. Radiocarbon dating uses the naturally occurring carbon-14 isotope as an isotopic label.
- Isotopes of a chemical element differ only in the mass number. For example, the isotopes of hydrogen can be written as ^{1}H, ^{2}H and ^{3}H, with the mass number superscripted to the left. When the atomic nucleus of an isotope is unstable, compounds containing this isotope are radioactive. Tritium is an example of a radioactive isotope.
- When a labelled chemical compound undergoes chemical reactions one or more of the products will contain the radioactive label. Analysis of what

happens to the radioactive isotope provides detailed information on the mechanism of the chemical reaction.

NOTES

4.10 KEY WORDS

- **Spectroscopy:** Spectroscopic studies were central to the development of quantum mechanics and included Max Planck's explanation of blackbody radiation, Albert Einstein's explanation of the photoelectric effect and Niels Bohr's explanation of atomic structure and spectra. Spectroscopy is used in physical and analytical chemistry because atoms and molecules have unique spectra.
- **Becquerel (Bq):** The Becquerel (Bq) is the SI derived unit of radioactivity. One Becquerel is defined as the activity of a quantity of radioactive material in which one nucleus decays per second.
- **Geiger Muller counter:** A Geiger counter is an instrument used for detecting and measuring ionising radiation. Also known as a Geiger–Müller counter (or Geiger–Muller counter), it is widely used in applications, such as radiation dosimetry, radiological protection, experimental physics, and the nuclear industry.
- **Radioactive tracer:** A radioactive tracer, radiotracer, or radioactive label, is a chemical compound in which one or more atoms have been replaced by a radionuclide so by virtue of its radioactive decay it can be used to explore the mechanism of chemical reactions by tracing the path that the radioisotope follows from reactants to products.
- **Autoradiograph:** An autoradiograph is an image on an x-ray film or nuclear emulsion produced by the pattern of decay emissions (e.g., beta particles or gamma rays) from a distribution of a radioactive substance.

4.11 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term spectroscopy.
2. Give the applications of spectroscopy.
3. Elaborate on the delayed effect of radiation.
4. How will you measure the radioactivity?
5. Explain the Geiger Muller counter.
6. Elaborate on the isotopes as tracers.
7. What is autoradiography?

Long-Answer Questions

1. Discuss briefly about the spectroscopy giving the principles and applications.
2. Describe the delayed effect of radiation.
3. Analyse the measurement of radioactivity.
4. Briefly explain about the Geiger Muller counter.
5. Explain in detail about the isotopes as tracers giving examples.
6. Discuss briefly about the autoradiography.

NOTES

4.12 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

BLOCK - II
BIOSTATISTICS

NOTES

UNIT 5 DEFINITION AND SCOPE OF BIOSTATISTICS

Structure

5.0 INTRODUCTION

Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It encompasses the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.

In this unit you will study about how statistics has become an integral part of our daily lives. Every day, we are confronted with some form of statistical information through different sources. Every raw data cannot be termed as statistics. Similarly, single or isolated facts or figures cannot be called statistics as these cannot be compared or related to other figures within the same framework. In this unit you will study that any quantitative and numerical data can be identified as statistics when it possesses certain identifiable characteristics as per the norms of statistics. The area of statistics can primarily be split up into two identifiable sub-areas termed as, Descriptive Statistics and Inferential Statistics. In this unit you will also study about some of the terms used extensively in the field of statistics for scientific measurement. Statistical investigation is a comprehensive process and requires systematic collection of data about some group of people or objects, describing and organising the data, analysing this data with the help of different statistical methods, summarising the analysis and using these results for making true judgements, decisions and predictions. The validity and accuracy of final judgement depends on the process and methodology of collection of primary

data. The quality of data greatly affects the conclusions and hence, utmost importance must be given to this process and every possible precaution should be taken to accurately collect the primary and secondary data.

In this unit, you will study about the definition and scope of biostatistics, collection of data, primary and secondary data.

NOTES

5.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the concept and scope of biostatistics
- Discuss about the process of data collection
- Differentiate between primary data and secondary data

5.2 DEFINITION AND SCOPE OF BIOSTATISTICS

Biostatistics (also known as **biometry**) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.

Biostatistics and Genetics

Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results. Some genetics scientists even contributed with statistical advances with the development of methods and tools. Gregor Mendel started the genetics studies investigating genetics segregation patterns in families of peas and used statistics to explain the collected data. In the early 1900s, after the rediscovery of Mendel's Mendelian inheritance work, there were gaps in understanding between genetics and evolutionary Darwinism. Francis Galton tried to expand Mendel's discoveries with human data and proposed a different model with fractions of the heredity coming from each ancestral composing an infinite series. He called this the theory of 'Law of Ancestral Heredity'. His ideas were strongly disagreed by William Bateson, who followed Mendel's conclusions that genetic inheritance were exclusively from the parents, half from each of them.

Definition

Most business decisions are made today on the basis of relevant information and statistical analysis of such information. Quantitative analysis has replaced intuition and experienced guess work in solving most business problems. One of the tools to understand information is statistics.

NOTES

In general, business statistics can be defined as 'a body of methods for obtaining, organizing, summarizing, presenting, interpreting, analysing and acting upon numerical facts related to an activity of interest. Numerical facts are usually subjected to statistical analysis with a view to helping a decision-maker make wise decisions in the face of uncertainty'.

The word 'statistics' can be referred to in two ways. In a common way, it refers simply to numerical statements of facts such as the number of children in a family, the number of books on statistics in the college library, the number of students enrolled in the department of economics in Delhi University, and so on. The following statements indicate the use of statistics as referring to numbers.

- Around 20 million Americans have a serious drinking problem.
- Nearly 52,000 Americans died in automobile accidents last year.
- More than 76 per cent voters turned out to vote during elections in Punjab in February 2007.
- Majority of Americans consider Japanese cars superior in quality than American cars.

All these statements represent statistical conclusions in some form. These conclusions help us in formulating specific policies and attitudes with respect to diverse areas of interest.

The second meaning of statistics refers to the field of study rather than simply to numerical statements. As an area of study, it is primarily concerned with making scientific and rational decisions about various properties and characteristics of some population of interest, such as stock market trends, interest rates, demographic shifts, inflation rates over the years, and so on. Consider the following statistical statements:

- The crime rate in the city has gone up by 15 per cent over what it was last year. (This statistical conclusion could help us in making decisions regarding our safety and security in the city).
- The rate of inflation is expected to remain less than 5 per cent per year over the next five years. (This could help us in making more educated judgements about the general economic health of the country in the near future).
- Less than 20 per cent of all high school graduates enter colleges for higher education and less than 40 per cent of those who do enter colleges actually graduate. (This statement gives us a good indication of the educational philosophy of the country and the community and the reasons for such low rates of admission into colleges and graduation could be investigated).

All these statements represent statistical conclusions in some form, which help us understand our environment better, and further help us in formulating specific policies and attitudes to address and solve issues of interest.

Descriptive Statistics

As the name suggests, descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data. These methods can either be graphical or computational. Thus data can be presented in the form of a chart or a table in order to show certain trends, proportions, maximum and minimum values, and so on. For example, if we simply describe the number of workers in different types of industries in America, then that would constitute descriptive statistics. In addition to the organization of data, the field of descriptive statistics is concerned with the analysis of data so that the data can be easily understood. Averages, proportions and other measures that describe the spread of data around the average are also some of the measures used to describe the data. By using these measures, we summarize the data and even though we may lose the detail, we gain clarity and compactness. For example, the following statistics, in their most summarized presentation describe in some way the characteristics of the population from which they were drawn.

- The ages of students in my statistics class range from 19 to 45 years.
- The average IQ of students at our college is 140.
- 20 per cent of the students in my class are married.

All these examples simply summarize and describe the data. Not much can be inferred from them, nor can definite decisions be made or conclusions drawn.

For a proper appreciation of the various descriptive statistics involved, it is necessary to note that most of the statistical distribution have some common features. Though the size of the variables varies from item to item, most of the items are distributed in such a manner that if we move from the lowest value to the highest value of the variable, the number of items at each successive stage increases with a certain amount of regularity till we reach a maximum; and then as we proceed further, they decrease with the similar regularity. If we plot the percentage frequency density, i.e., the percentage of cases in an interval of unit variable width, we get frequency curves of the type shown in Fig. 5.1. (Note that the area under each curve should be equal to 100, the total percentage points).

There are various 'gross' ways in which frequency curves can differ from one another. Even when the 'general' shapes of the curves are the same (the area under them already made equal by the strategy of plotting the per cent density), the details of the shape may change. Thus the curve *B* has a smaller spread than *A*, the curve *C* is more peaky and curve *E* is less symmetrical. Even when the curves have almost the same shape (i.e., same spread, peakness, symmetry, etc.) as in curves *A* and *D*, the two may differ in location along the variable axis. Thus the items of distribution *D* are generally larger than those of *A*. So are those of *B* compared to *A*. Thus, a kind of an 'average' location of the distribution along the

NOTES

variable axis is an important descriptive statistics. These statistics are collectively known as measures of location or of central tendency.

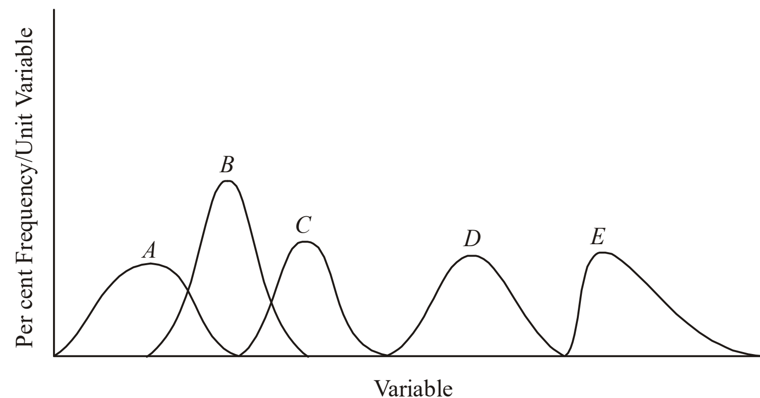


Fig. 5.1 *Representation of Measures of Central Tendency*

Inferential Statistics

Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population. The measured characteristics of the sample are known as sample statistics, while the measured characteristics of the population are known as population parameters. A major portion of statistics deals with making decisions, inferences, predictions and forecasts about the population based on the results obtained from samples taken from such populations.

The need for inferential statistical methods derives from the need for sampling. As the population becomes large, it is usually too costly, too time consuming and too cumbersome to take the entire population into consideration in order to obtain our information of interest. Of course, the results obtained from the entire population are the most accurate and if the population indeed is small, then it is advisable to consider the entire population. However, when the population is large, sometimes considered infinite, then sampling method is used.

The question is, How do these sample statistics relate to population parameters? Can we state that the conclusions drawn from the analysis of the sample are exactly the same as the conclusions that would be drawn from the entire population from which the representative sample was taken? The answer is unlikely. How close is the sample characteristics to the population characteristics would depend upon the randomness of the sample as well as the size of the sample. The more random the sample is and larger the sample is, the more closely its characteristics would be with the population characteristics. This link, in terms of the degree of closeness is provided by probability theory. Probability theory provides the link by ascertaining the likelihood that the results from the sample reflect the results from the population.

Our interest is not in finding the characteristics of a sample but our to find the characteristics of the population. Sampling is simply a means to the end. For example, if we want to know the salary of university professors, we mean the

salary of all university professors and not simply of the sample we have taken. Only then can observations and decisions be made in this regard. Similarly, if we want to know what percentage of eligible voters will vote for Congress in the next general elections in India, a sample in itself would not indicate that, and we cannot ask the entire population. Our decisions and projections would be based on the inclination of the entire population. A sample in itself would not mean much, if any thing. How ever, if the sample truly represents the population, then we can draw conclusions about the population on the basis of sample results. Appended to these conclusions will be a probability statement specifying the likelihood or confidence that the results from the sample reflect the voting behaviour of the population. Usually, the margin of error is stated as plus or minus three to five per cent.

Statistical inference deals with methods of inferring or drawing conclusions about the characteristics of the population based upon the results of the sample taken from the same population. The measured characteristics of the sample are called sample statistics and the measured characteristics of the population are known as population parameters. The question is: How do these sample statistics relate to population parameters? Can we state that the conclusions drawn from the analysis of the sample are exactly the same as the conclusions that would be drawn from the entire population from which the representative sample was taken?

Following are some of the situations that the field of inferential statistics deals with.

(a) Between 35 per cent and 40 per cent of graduate students in the universities are married. These statistics refer to the entire population of graduate students. It would be reasonable to assume that these percentages were calculated on the basis of samples taken from the population of all graduate students. The students in these samples were asked in order to know as to how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of the graduate students.

(b) There is a definitive association between smoking and lung cancer. This statement is the result of endless research on many samples taken and studied in order to find out if there was any correlation between smoking and lung cancer and based upon the results thus obtained from sample studies, a valid statement about the association of smoking with lung cancer in the whole population can be made.

(c) 30 per cent of all television viewers watched the show 20/20 last night. This statement can be compared with the following statement: 30 per cent of those who were interviewed watched the show 20/20 last night. The latter statement is descriptive statistics since it is only presenting the data in a summarized form. However, if we infer from the second statement to reach at the first statement, then the first statement is an example of statistical inference.

(d) Suppose that the Chancellor of Punjabi University wanted to conduct a survey to learn about student perceptions concerning the quality of life on campus. The population will be all the students enrolled in the

NOTES

university, while a sample will consist of only the students who have been randomly selected to be included in the sample to participate in the survey. The goal is to determine various attitudes and characteristics of interest relating to quality of student life in the entire university using the sample statistics to draw conclusions about the similar population characteristics

(e) Between 35 per cent to 40 per cent of graduate students in the universities are married. These statistics refer to the entire population of graduate students. It would be reasonable to presume that these percentages were calculated on the basis of samples taken from the population of all graduate students. The students in these samples were asked in order to know how many of these students were married. The answers formed the basis for drawing conclusions about the entire population of graduate students.

(f) There is a definite association between smoking and lung cancer. This statement is the result of endless research on many samples taken and studied in order to find out if there was any correlation between smoking and lung cancer and based upon the results thus obtained from these sample studies, a valid statement about the association of smoking with lung cancer in the whole population could be made.

5.3 COLLECTION OF DATA

Data

Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense, such as a set of data.) For example, if we ask the students in a classroom their ages and we write down their ages as they tell us, then a collection of these numbers would be considered as data. Similarly, information regarding incomes of families, IQ scores of students, examination result scores of students in a class, heights of policemen in New York city, and so on, when collected, is known as data. If this data is written down as collected, then it is known as raw data. If this data is written in ascending or descending order, then it would be called ordered data. If this ordered data is arranged in arrays of rows and columns, then the data is known to be presented in an ordered array.

Variable

A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people. For example, Mr X is a variable since X can represent anybody. On the other hand, a constant will always have the same value. For example, the number of days in a week are constant and will always remain the same. Consider the following illustration:

Let, $x + 6 > 10$ be an inequality. Now, if x is a whole number, then it can have any value greater than 4. While the values 6 and 10 are constant and do not

change, x can be 5, 6, 7... and up to any value. Thus x is a variable which can have any number of different values.

There are two types of variables. One type is known as discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library. Similarly, the number of children in a family would be considered as values of a discrete variable, since the children can be counted exactly.

On the other hand, a continuous variable is a variable whose values can theoretically take on an infinite number of values within a given range of values.

Hence, these values are measured as against being counted. However, since the measurement value would depend upon how accurately we measure it, any exact value would simply be one of the infinite number of values on a continuous scale between two given points. For example, the height of a child touches every one of the infinite number of points between, let us say, 40 inches and 40.1 inches as he/she grows from 40 inches to 40.1 inches. Accordingly, the value of a continuous variable is more accurately defined if it is stated as being between two points such as 40 inches and 40.1 inches.

A random variable

A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response. By definition of randomness, each possible entity has the same chance of being considered. For instance, lottery drawings are considered to be random drawings so that each number has exactly the same chance of being picked up. Similarly, the value of the outcome of a toss of a fair coin is random, since a head or a tail has the same chance of occurring.

A random variable may be qualitative or quantitative in nature. The qualitative random variables yield categorical responses so that the responses fit into one category or another. For example, a response to a question such as 'Are you currently unemployed?' would fit in the category of either 'yes' or 'no'. On the other hand, quantitative random variables yield numerical responses. For example, responses to questions such as, 'How many rooms are there in your house?' or 'How many children are there in the family?' would be in numerical values. Also, these values being whole numbers are considered discrete values. These are the values of discrete quantitative random variables. On the other hand, responses to questions like, 'How tall are you?' or 'How much do you weigh?' would be values of continuous quantitative random variables, since these values are measured and not counted. Some of the examples of these variables are:

(a) ***Qualitative random variables***

- Sex of students in the class
- Political affiliation of a faculty member in the college
- Opinions of economists regarding the economic conditions in the country

NOTES

(b) ***Quantitative random variables***

(*i*) Discrete quantitative random variables

- Number of people attending a conference
- Number of eggs in the refrigerator
- Number of children at a summer camp

(*ii*) Continuous quantitative random variables

- Heights of models in a beauty contest
- Weights of people joining a diet programme
- Lengths of steel bars produced in a given production run

Sample

A sample is a portion of the total population that is considered for study and analysis. For instance, if we want to study the income pattern of professors at City University of New York and there are 10,000 professors, then we may take a random sample of only 1,000 professors out of this entire population of 10,000 for the purpose of our study. Then this number of 1,000 professors constitutes a sample. The summary measure that describes a characteristic such as average income of this sample is known as a statistic.

Sampling is the process of selecting a sample from the population. It is technically and economically not feasible to take the entire population for analysis. So we must take a representative sample out of this population for the purpose of such analysis. A sample is part of the whole, selected in such a manner as to be representing the whole.

A random sample

It is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the sample taken from the population would be truly representative of the population. The degree of randomness of selection would depend upon the process of selecting the items from the sample. A true random sample would be free from all biases whatsoever. For example, if we want to take a random sample of five students from a class of twenty-five students, then each one of these twenty-five students should have the same chance of being selected into the sample. One way to do this would be writing the names of all students on separate but small pieces of paper, folding each piece of this paper in a similar manner, putting each folded piece into a container, mixing them thoroughly and drawing out five pieces of paper from this container.

Sampling without replacement

The sample as taken in the above example is known as sampling without replacement, because each person can only be selected once so that once a piece

of paper is taken out of the container, it is kept aside so that the person whose name appears on this piece of paper has no chance of being selected again.

NOTES

Sampling with replacement

There are certain situations in which the piece of paper once selected and taken into consideration is put back into the container in such a manner that the same person has the same chance of being selected again as any other person. For example, if we are randomly selecting five persons for award of prizes so that each person is eligible for any and all prizes, then once the slip of paper is drawn out of the container and the prize is awarded to the person whose name appears on the paper, the same piece of paper is put back into the container and the same person has the same chance of winning the second prize as anybody else.

Random Number Tables

For a sample to be truly representative of the population, it must truly be random. To make the random selection easier, we can make use of tables of random numbers which are generated by computers. A perfect random number table would be one in which every digit has been entered randomly. This means that no matter where you start within the table and no matter in which direction you move, the probability of encountering any one of the ten digits (0, 1, 2,...9) would be the same. This means that the chance of any one of these digits being at any place in the table is exactly one out of ten. Similarly, if these digits are grouped in pairs (00, 01, 02,...99), then each of these pairs have the same chance of occurring at any place so that each pair would have a chance of occurring of one out of a hundred.

The following is an example of a random number Table 5.1:

Fig. 5.1 *Random Number Table*

Column Number

Row Number	1	2	3	4	5
1	81625	42372	07090	23422	10742
2	20891	27833	93079	16274	92818
3	62882	48722	39630	96434	09895
4	59882	84713	82521	29026	08591
5	17932	14360	42933	89380	68191
6	67732	36772	09281	26898	30919
7	58198	87824	47958	04701	17369
8	57041	47778	02361	86939	61463
9	05264	49678	02067	58121	61822
10	84935	60407	16547	21359	58913

As an example of use of random number tables, let us assume that we have to select a random sample from a finite population. The population cannot be infinite due to the limitation as to how far the random numbers can go. Let there be 100 students in the population from which we have to draw a sample of five students. Now we assign a two-digit number to each member of the population so that each member is known as 00, 01, 02 ... 99. For selecting five students at random from this population, we go to the random number table with groups of two digits each and starting at any point and moving in any direction we pick the five groups of numbers. Suppose that the numbers picked up are 07, 22, 23, 58 and 78. Then those members of the population to whom these numbers are assigned constitute the random sample. In case we want to use a random number table in which groups of five digits are arranged, as in the table, then we can use only the first two digits or any two digits out of the five and reach the same conclusion of randomness. In this given table suppose we pick row five and go across and pick up the first two digits from each group of five, we get the following numbers: 17, 14, 42, 89 and 68. Thus, those five members of the population to whom these numbers are assigned constitute the random sample.

Sources of Data and Methods of Data Collection

The following are some of the sources of data to collect first hand information.

- Census
- World Bank
- WHO (World Health Organization)
- NSSO (National Sample Survey Organization)
- Economic Survey
- Civil Registration
- Sample Registration System
- National Family and Health Surveys
- Reproductive and Child Health Project
- SRS Surveys
- Multiple Indicator Survey
- Medical Causes of Death
- Demographic and Health Surveys

Since the quality of the results obtained from statistical data for the purpose of using these outcomes for managerial decision-making depends upon the quality of the information itself collected, it is important that a sound investigative process be established to ensure that the data are highly representative and highly unbiased. This requires a high degree of skill and also certain precautionary measures are to be taken.

The following steps may be considered in the primary data collection process:

Planning the study

Before any procedures for data collection are established, the purpose and the scope of the study must be clearly specified. If any similar studies have been

conducted, prior to the current one, then the investigator may want to use some secondary data in his own study, and may redefine his objectives on the basis of the previous studies conducted. The scope of the study must take into consideration the field to be covered, and the time period in which to conduct the study. The time span is very important, because in certain areas, the conditions change very quickly, and hence by the time the study is completed, it may become irrelevant. The statistical units and the desired accuracy of such units must be clearly specified.

Methods of Collecting Primary Data

Primary data can be collected by anyone or more of the following methods:

(a) **Direct Personal Observation.** Under this method, the investigator presents himself personally before the informant and obtains a first hand information. This method is most suitable when the field of enquiry is small and a greater degree of accuracy is required.

Merits

(*i*) The first hand information obtained by the investigator is bound to be more reliable and accurate since the investigator can extract the correct information by removing doubts, if any, in the minds of the respondents regarding certain questions.

(*ii*) High response rate since the answers to various questions are obtained on the spot.

(*iii*) It permits explanation of questions concerning difficult subject matter.

(*iv*) It permits evaluation of respondent, his circumstances and reliability.

(*v*) This method is useful where sponteneity of response is required.

(*vi*) It provides personal rapport which helps to overcome reluctance to respond.

(*vii*) Where the investigator and informant talk face to face, it becomes possible to explore questions in depth.

(*viii*) Information is collected promptly and there is no dribbling in.

Limitations

(*i*) This method is suitable only for intensive studies and not for extensive enquiries.

(*ii*) This method is time-consuming and the investigation may have to be spanned over a long period.

(*iii*) This method is highly subjective in nature and the results of the enquiry may be adversely affected by the personal biases, whim and prejudices of the investigator.

(b) **Telephone Survey.** Under this method, the investigator, instead of presenting himself before the informants, contacts them on telephone and collects information from them.

NOTES

Merits

(*i*) The method is more convenient than personal interview.

(*ii*) This method is less time-consuming and can be applied even to extensive fields of enquiries. Telephone survey has all the other merits of personal interview.

Limitations

(*i*) This method excludes those who do not have a telephone as also those who have unlisted telephones.

(*ii*) This method is also subjective in nature and personal bias, whim and prejudices of the investigator may adversely affect the results of the enquiry.

(c) **Indirect Personal Interview.** Under this method, instead of directly approaching the informants, the investigator interviews several third persons who are directly or indirectly concerned with the subject matter of the enquiry and who are in possession of the requisite information. Such a procedure is followed by the enquiry committees and commissions appointed by the Government of India. The committee selects persons known as witnesses and collects information from them by getting answers to questions decided in advance. This method is highly suitable where the direct personal investigation is not practicable either because the informants are unwilling or reluctant to supply the information or where the information desired is complex and the study in hand is extensive.

Merits

(*i*) This method is less costly and less time-consuming than the direct personal investigation.

(*ii*) Under this method, the enquiry can be formulated and conducted more effectively and efficiently as it is possible to obtain the views and suggestions of the experts on the given problem.

Limitations

The success of this method depends upon:

(*i*) The representative character of the witnesses

(*ii*) The personal knowledge of the witnesses about the subject matter of enquiry

(*iii*) The personal prejudices of the witnesses as regards definiteness in stating what is wanted

(*iv*) The ability of the interviewer to extract information from the witnesses by asking appropriate questions and cross-questions

(d) **Information Received Through Local Agents.** Under this method, the information is not collected formally by the investigator, but local agents, commonly known as correspondents, are appointed in different parts of the area under investigation. These agents collect information in their areas and transmit the same to the investigator. They apply their own judgement as to the best method of obtaining information. This method is usually employed by newspaper or periodical agencies which require information in different fields such as economic trends, business, stock and share markets, sports, politics, and so on.

Merits

(*i*) This method is very cheap and economical for extensive investigations.

(*ii*) The required information can be obtained expeditiously since only rough estimates are required.

Limitations

Since the correspondents apply their own judgement about the method of collecting the information, the results are often vitiated due to personal prejudices and whims of the correspondents. The data so obtained is thus not so reliable. This method is suitable only if the purpose of investigation is to obtain rough and approximate estimates. It is unsuited where a high degree of accuracy is desired.

(e) **Mailed Questionnaire Method.** Under this method, the investigator prepares a questionnaire containing a number of questions pertaining to the field of enquiry. These questionnaires are sent by post to the informants together with a polite covering letter explaining in detail the aims and objectives of collecting the information, and requesting the respondents to cooperate by furnishing the correct replies and returning the questionnaire duly filled in. In order to ensure quick response, the return postage expenses are usually borne by the investigator. This method is usually adopted by the research workers, private individuals and non-official agencies. The success of this method depends upon the proper drafting of the questionnaire and the cooperation of the respondents.

Merits

(*i*) By this method, a large field of investigation may be covered at a very low cost. In fact, this is the most economical method in terms of time, money and manpower.

(*ii*) Errors due to personal bias of the investigators or enumerators are completely eliminated as the information is supplied by the person concerned in his own handwriting.

Limitations

(*i*) This method can be used only if the respondents are educated and can understand the questions well, and reply in their own handwriting.

(*ii*) Sometimes, the informants may not send back the schedules and even if they return the schedules, they may be incorrectly filled in.

(*iii*) Sometimes, the informants are not willing to give written information in their own handwriting on certain personal questions like income, personal habits and property.

(*iv*) There is no scope for asking supplementary questions for cross-checking of the information supplied by the respondents.

(f) **Questionnaire Sent through Enumerators.** Under this method, instead of sending the questionnaire through post, the investigator appoints agents known as enumerators, who go to the respondents personally with the questionnaire, ask them the questions given therein, and record their replies. This method is generally used by business houses, large public enterprises and research institutions.

Merits

(*i*) The information collected through this method is more reliable as the enumerators can explain in detail the objectives and aims of the enquiry to the respondents and win their cooperation.

(*ii*) Since the enumerators personally call on the respondents, there is very little non-response.

(*iii*) This technique can be used with advantage even if the respondents are illiterate.

(*iv*) The enumerators can effectively check the accuracy of the information supplied through some intelligent cross-questioning by asking supplementary questions.

Limitations

(*i*) The method is more expensive and can be used by financially strong bodies or institutions only.

(*ii*) It is more time-consuming than the mailed questionnaire method.

(*iii*) The success of the method depends upon the skill and efficiency of the enumerators to collect the information as also on the efficiency and wisdom with which the questionnaire is drafted.

Sources and Methods of Collecting Secondary Data

The chief sources of secondary data may be broadly classified into the following two groups:

(*i*) Published sources

(*ii*) Unpublished sources

(*i*) **Published sources:** There are a number of national organizations and international agencies which collect and publish statistical data relating to business, trade, labour, price, consumption, production, etc. These publications are useful sources of secondary data. Some of these published sources are as follows:

1. Official publications of the Central and State Governments such as monthly abstract of statistics, national income statistics and vital statistics of India.
2. Publications of semi-government organizations, e.g., the Reserve Bank of India bulletin.
3. Publications of research institutions, e.g., the publications of the Indian Council of Agricultural Research (I.C.A.R.), New Delhi.
4. Publications of commercial and financial institutions, e.g., the publications of the F.I.C.C.I.
5. Reports of various committees and commissions appointed by the government, such as the Wanchoo Commission Report on Taxation.
6. Newspapers and periodicals like *Economic Times, Statesman Year Book* also publish useful statistical data.
7. International publications like the *U.N. Statistical Year Book, Demographic Year Book,* etc.

(*ii*) **Unpublished sources:** The records maintained by private firms or business houses which may not like to release their data to any outside agency; the research carried out by the research scholars in the universities or research institutes may also provide useful statistical data.

Precautions in the use of secondary data: Secondary data should be used with extra caution since they have been collected by someone other than the investigator. Before using such data the investigator must be satisfied in regard to the reliability, accuracy, adequacy and suitability of the data to the given problem under investigation. Before using secondary data, the investigator should examine the following questions.

1. Is the data suitable for the purpose of investigation? For this, he should compare the objectives, nature and scope of the given enquiry with the original investigation. He should also confirm that the various terms and units were clearly defined and were uniform throughout the earlier investigation and these definitions are suitable for the present enquiry as well.
2. Is the data reliable? For this, the investigator himself should satisfy about (*i*) the reliability, integrity and experience of the collecting organization, (*ii*) the reliability of the source of information, (*iii*) the methods used for the collection and analysis of the data, and (*iv*) the degree of accuracy desired by the company.

NOTES

3. Is the data adequate? Adequacy of data is to be judged in the light of the requirements of the survey and the geographical areas covered by the available data. Adequacy of data is also to be considered in the light of the time period for which the data is available.

Hence, in order to arrive at conclusions free from limitations and inaccuracies, the secondary data must be subjected to thorough scrutiny and editing before it is accepted for use.

Sample Selection

The third step in the primary data collection process is selecting an adequate sample. It is necessary to take a representative sample from the population, since it is extremely costly, time-consuming and cumbersome to do a complete census. Then, depending upon the conclusions drawn from the study of the characteristics of such a sample, we can draw inferences about the similar characteristics of the population. If the sample is truly representative of the population, then the characteristics of the sample can be considered to be the same as those of the entire population. For example, the taste of soup in the entire pot of soup can be determined by tasting one spoonful from the pot if the soup is well stirred. Similarly, a small amount of blood sample taken from a patient can determine whether the patient's sugar level is normal or not. This is so because the small sample of blood is truly representative of the entire blood supply in the body.

Sampling is necessary because of the following reasons: First, as discussed earlier, it is not technically or economically feasible to take the entire population into consideration. Second, due to dynamic changes in business, industrial and social environment, it is necessary to make quick decisions based upon the analysis of information. Managers seldom have the time to collect and process data for the entire population. Thus, a sample is necessary to save time. The time element has further importance in that if the data collection takes a long time, then the values of some characteristics may change over the period of time so that data may no longer be up to date, thus defeating the very purpose of data analysis. Third, samples, if representative, may yield more accurate results than the total census. This is due to the fact that samples can be more accurately supervised and data can be more carefully selected. Additionally, because of the smaller size of the samples, the routine errors that are introduced in the sampling process can be kept at a minimum. Fourth, the quality of some products must be tested by destroying the products. For example, in testing cars for their ability to withstand accidents at various speeds, the environment of accidents must be simulated. Thus, a sample of cars must be selected and subjected to accidents by remote control. Naturally, the entire population of cars cannot be subjected to these accident tests and hence, a sample must be selected.

One important aspect to be considered is the size of the sample. The sampling size—which is the number of sampling units selected from the population for investigation—must be optimum. If the sample size is too small, it may not

appropriately represent the population or the universe as it is known, thus leading to incorrect inferences. Too large a sample would be costly in terms of time and money. The optimum sample size should fulfil the requirements of efficiency, representativeness, reliability and flexibility. What is an optimum sample size is also open to question. Some experts have suggested that 5 per cent of the population properly selected would constitute an adequate sample, while others have suggested as high as 10 per cent depending upon the size of the population under study. However, proper selection and representation of the sample is more important than size itself. The following considerations may be taken into account in deciding about the sample size:

(a) The larger the size of the population, the larger should be the sample size.

(b) If the resources available do not put a heavy constraint on the sample size, a larger sample would be desirable.

(c) If the samples are selected by scientific methods, a larger sample size would ensure greater degree of accuracy in conclusions.

(d) A smaller sample could adequately represent the population, if the population consists of mostly homogeneous units. A heterogeneous universe would require a larger sample.

Editing the Primary Data

Once a set of data has been collected, it is necessary to process it for proper presentation. Editing of data is required as preparatory work before tabulation and statistical analysis is carried out. The editing process would be required to ensure that the data is complete and as required. In the case of the questionnaire method of gathering data, it should be made certain that all the questions have been answered. Additionally, responses are scrutinized to make sure that there are no contradictions among different answers of the same respondent. If so, then the respondent can be contacted again to clarify such contradictions. Editing would also help eliminate inconsistencies or obvious errors due to arithmetical treatment.

When the data is to be processed by computers, then it must be coded and converted into the computer language. For some qualitative characteristics, code numbers can be assigned and identified. For instance, the response to a question such as 'Are you married or single?', a code of digit 1 can be assigned to the qualitative answer 'married' and a code of 0 to the answer 'single'. This coding job should be done while editing the data.

5.4 PRIMARY AND SECONDARY DATA

The statistical data, as previously discussed, may be classified under two categories depending upon the sources utilized. These categories are:

1. **Primary Data.** Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research

institutions. For example, if a researcher is interested to know what the women think about the issue of abortion, he/she must undertake a survey and collect data on the opinions of women by asking relevant questions. Such data collected would be considered as primary data.

2. **Secondary Data.** When an investigator uses the data which has already been collected by others, such data is called secondary data. This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes. The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on. For example, if a researcher desires to analyse the weather conditions of different regions, he can get the required information or data from the records of the meteorology department. Even though secondary data is less expensive to collect in terms of money and time, the quality of this data may even be better under certain situations because it may have been collected by persons who were specifically trained for that purpose. However, such secondary data must be used with the utmost care. The reason is that such data may be full of errors due to the fact that the purpose of the collection of data by the primary agency may have been different than that of the user of the secondary data. Additionally, there may have been biases introduced in collection of data or analysis of data. For example, the size of the sample may have been inadequate or there may have been arithmetical or definitional errors. Hence, it is necessary to critically investigate the validity of the secondary data as well as the credibility of the primary data collection agency.

When the raw data has been collected and edited, it should be put into an ordered form (ascending or descending order) so that it can be looked at more objectively. The next important step towards processing the data is classification. Classification means separating items according to similar characteristics and grouping them into various classes. The items in different classes will differ from each other on the basis of some characteristics or attributes. Classification of data is very similar to sorting of mail at a post office, where mail is classified according to its geographical destination and may further be classified into the type of mail such as first class, parcel post, and so on. The data may be classified into four broad classes:

(a) **Geographical.** This classification groups the data according to locational differences among the items. The geographical areas are usually listed in alphabetical order for easy reference. For example, the book listing the colleges and universities in various states in America would first list the states in an alphabetical order and then the colleges and universities within these states in an alphabetical order.

(b) **Chronological.** Chronological classification includes data according to the time period in which the items under consideration occurred.

NOTES

For example, the sales of automobiles in America over the last ten years may be grouped according to the year in which such sales took place.

(c) Qualitative. In this type of classification, the data is grouped together according to some distinguished characteristic or attribute such as religion, sex, age, national origin, and so on. This classification simply identifies whether a given attribute is present or absent in a given population. For example, the population may be divided into two classes of males and females. Then the attribute of male will go into one class and attribute of female will go into the other.

(d) Quantitative. It refers to the classification of data according to some attribute which has magnitude and can be measured such as classification according to weight, height, income and so on. For example, the salaries of professors at a university may be classified according to their rank of instructor, assistant professor, associate professor and full professor.

Check Your Progress

1. Give the definition of biostatistics.
2. Define the term statistics.
3. What is inferential statistics?
4. Explain about the data.
5. Elaborate on the primary data.
6. Interpret the secondary data.

5.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.
2. Most business decisions are made today on the basis of relevant information and statistical analysis of such information. Quantitative analysis has replaced intuition and experienced guess work in solving most business problems. One of the tools to understand information is statistics.
3. Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population.

4. Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense, such as a set of data.)
5. Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.
6. When an investigator uses the data which has already been collected by others, such data is called secondary data. This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes. The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on.

5.6 SUMMARY

- Biostatistical modelling forms an important part of numerous modern biological theories. Genetics studies, since its beginning, used statistical concepts to understand observed experimental results.
- In general, business statistics can be defined as 'a body of methods for obtaining, organizing, summarizing, presenting, interpreting, analysing and acting upon numerical facts related to an activity of interest. Numerical facts are usually subjected to statistical analysis with a view to helping a decision-maker make wise decisions in the face of uncertainty'.
- Descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data.
- Inferential statistics can be defined as those methods that are used to estimate a characteristic of a population or making of a decision concerning a population on the basis of the results obtained from a sample taken from the same population.
- The need for inferential statistical methods derives from the need for sampling. As the population becomes large, it is usually too costly, too time consuming and too cumbersome to take the entire population into consideration in order to obtain our information of interest.
- If this data is written down as collected, then it is known as raw data. If this data is written in ascending or descending order, then it would be called ordered data. If this ordered data is arranged in arrays of rows and columns, then the data is known to be presented in an ordered array.
- A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.

- Discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library. Similarly, the number of children in a family would be considered as values of a discrete variable, since the children can be counted exactly.
- A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response.
- A sample is a portion of the total population that is considered for study and analysis.
- Random sample is a collection of items selected from the population in such a manner that each item in the population has exactly the same chance of being selected, so that the sample taken from the population would be truly representative of the population.
- Direct Personal Observation method the investigator presents himself personally before the informant and obtains a first-hand information. This method is most suitable when the field of enquiry is small and a greater degree of accuracy is required.
- Indirect Personal Interview method, instead of directly approaching the informants, the investigator interviews several third persons who are directly or indirectly concerned with the subject matter of the enquiry and who are in possession of the requisite information.
- Unpublished sources records maintained by private firms or business houses which may not like to release their data to any outside agency; the research carried out by the research scholars in the universities or research institutes may also provide useful statistical data.
- Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.
- When an investigator uses the data which has already been collected by others, such data is called secondary data. This data is primary data for the agency that collected it and becomes secondary data for someone else who uses this data for his own purposes. The secondary data can be obtained from journals, reports, government publications, publications of professional and research organizations, and so on.
- Chronological classification includes data according to the time period in which the items under consideration occurred.
- Qualitative is the type of classification, the data is grouped together according to some distinguished characteristic or attribute, such as religion, sex, age, national origin, and so on. This classification simply identifies whether a given attribute is present or absent in a given population.

- Quantitative is the refers to the classification of data according to some attribute which has magnitude and can be measured, such as classification according to weight, height, income and so on.

NOTES

5.7 KEY WORDS

- **Biostatistics:** Biostatistics (also known as biometry) are the development and application of statistical methods to a wide range of topics in biology. It is includes the design of biological experiments, the collection and analysis of data from those experiments and the interpretation of the results.
- **Descriptive statistics:** Descriptive statistics merely describe the data and consist of methods and techniques used in collection, organization, presentation and analysis of data in order to describe the various features and characteristics of such data.
- **Data:** Data is simply the numerical results of any scientific measurement. (Data can also be used in singular sense such as a set of data.)
- **Variable:** A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.
- **Primary data:** Primary data is one which is collected by the investigator himself for the purpose of a specific inquiry or study. Such data is original in character and is generated by surveys conducted by individuals or research institutions.

5.8 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the concept of biostatistics.
2. Define the term statistics.
3. Elaborate on the descriptive statistics.
4. What do you understand by inferential statistics?
5. Elaborate on the data with example.
6. What is variable?
7. Explain the sample selection.
8. Distinguish between primary and secondary data.

NOTES

Long-Answer Questions

1. Discuss briefly the concept of biostatistics with the help of examples.
2. Analyse the collection of data with various types of methods.
3. What is sample? Describe the random sample and random number table.
4. Explain the primary and secondary data. Give appropriate examples.

5.9 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 6 TYPES OF SAMPLING

NOTES

Structure

6.0 INTRODUCTION

In statistics, quality assurance, and survey methodology, sampling is the selection of a subset (a statistical sample) of individuals from within a statistical population to estimate characteristics of the whole population. Statisticians attempt for the samples to represent the population in question. Two advantages of sampling are lower cost and faster data collection than measuring the entire population.

In statistics, a Simple Random Sample (or SRS) is a subset of individuals (a sample) chosen from a larger set (a population) in which a subset of individuals are chosen randomly, all with the same probability. A simple random sample is an unbiased surveying technique. Simple random sampling is a basic type of sampling and can be a component of other more complex sampling methods.

In statistics, stratified sampling is a method of sampling from a population which can be partitioned into subpopulations. Stratification is the process of dividing members of the population into homogeneous subgroups before sampling. The strata should define a partition of the population.

A variable is any characteristics, number, or quantity that can be measured or counted. A variable may also be called a data item. Age, sex, business income and expenses, country of birth, capital expenditure, class grades, and eye colour and vehicle type are examples of variables.

In this unit, you will study about the types of sampling, random sampling, stratified random sampling, types of variables.

NOTES

6.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the types of sampling
- Define the random sampling
- Analyse the stratified random sampling
- Discuss about the types of variables

6.2 TYPES OF SAMPLING

Under census or complete enumeration survey method, data is collected for each and every unit (e.g., person, consumer, employee, household, organization) of the population or universe which are the complete set of entities and which are of interest in any particular situation. In spite of the benefits of such an all-inclusive approach, it is infeasible in most of the situations. Besides the time and resource constraints of the researcher, infinite or huge population, the incidental destruction of the population unit during the evaluation process (as in the case of bullets, explosives, etc), cases of data obsolescence (by the time census ends) do not permit this mode of data collection.

Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population. The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population. It should be clear to the researcher that a sample is studied not for its own sake, but the basic objective of its study is to draw inference about the population. In other words, sampling is a tool which helps us know the characteristics of the universe or the population by examining only a small part of it. The values obtained from the study of a sample, such as the average and dispersion are known as 'statistics' and the corresponding such values for the population are called 'parameters'.

Although diversity is a universal quality of mass data, every population has characteristic properties with limited variation. The following two laws of statistics are very important in this regard.

1. The law of statistical regularity states that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group. By random selection, we mean a selection where each and every item of the population has an equal chance of being selected.

NOTES

2. The law of inertia of large numbers states that, other things being equal, larger the size of the sample, more accurate the results are likely to be.

Hence, a sound sampling procedure should result in a representative, adequate and homogeneous sample while ensuring that the selection of items should occur independently of one another.

6.3 RANDOM SAMPLING

It refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term 'arbitrary' for 'random'. To ensure randomness, one may adopt either the lottery method or consult the table of random numbers, preferably the latter. Being a random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample. Furthermore, it is easy to assess the accuracy of the sampling estimates because sampling errors follow the principles of chance. However, a completely catalogued universe is a prerequisite for this method. The sample size requirements would be usually larger under random sampling than under stratified random sampling, to ensure statistical reliability. It may escalate the cost of collecting data as the cases selected by random sampling tend to be too widely dispersed geographically.

6.4 STRATIFIED RANDOM SAMPLING

In this method, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive but collectively exhaustive based on a variable known to be correlated with the variable of interest. Then, a simple random sample is chosen independently from each group. This method differs from simple random sampling in that in the latter the sample items are chosen at random from the entire universe. In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum. If the ratio of items between various strata in the population matches with the ratio of corresponding items between various strata in the sample, it is called proportionate stratified sampling; otherwise, it is known as disproportionate stratified sampling. Ideally, we should assign greater representation to a stratum with a larger dispersion and smaller representation to one with small variation. Hence, it results in a more representative sample than simple random sampling.

6.5 TYPES OF VARIABLES

A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people. For example, Mr *X* is a variable since *X* can represent anybody.

NOTES

On the other hand, a constant will always have the same value. For example, the number of days in a week are constant and will always remain the same. Consider the following illustration:

Let, $x + 6 > 10$ be an inequality. Now, if x is a whole number, then it can have any value greater than 4. While the values 6 and 10 are constant and do not change, x can be 5, 6, 7... and up to any value. Thus x is a variable which can have any number of different values.

There are two types of variables. One type is known as discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library. Similarly, the number of children in a family would be considered as values of a discrete variable, since the children can be counted exactly.

On the other hand, a continuous variable is a variable whose values can theoretically take on an infinite number of values within a given range of values.

Hence, these values are measured as against being counted. However, since the measurement value would depend upon how accurately we measure it, any exact value would simply be one of the infinite number of values on a continuous scale between two given points. For example, the height of a child touches every one of the infinite number of points between, let us say, 40 inches and 40.1 inches as he/she grows from 40 inches to 40.1 inches. Accordingly, the value of a continuous variable is more accurately defined if it is stated as being between two points such as 40 inches and 40.1 inches.

A random variable

A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response. By definition of randomness, each possible entity has the same chance of being considered. For instance, lottery drawings are considered to be random drawings so that each number has exactly the same chance of being picked up. Similarly, the value of the outcome of a toss of a fair coin is random, since a head or a tail has the same chance of occurring.

A random variable may be qualitative or quantitative in nature. The qualitative random variables yield categorical responses so that the responses fit into one category or another. For example, a response to a question such as 'Are you currently unemployed?' would fit in the category of either 'yes' or 'no'. On the other hand, quantitative random variables yield numerical responses. For example, responses to questions such as, 'How many rooms are there in your house?' or 'How many children are there in the family?' would be in numerical values. Also, these values being whole numbers are considered discrete values. These are the values of discrete quantitative random variables. On the other hand, responses to questions like, 'How tall are you?' or 'How much do you weigh?' would be values

NOTES

of continuous quantitative random variables, since these values are measured and not counted. Some of the examples of these variables are:

(a) *Qualitative random variables*

- Sex of students in the class
- Political affiliation of a faculty member in the college
- Opinions of economists regarding the economic conditions in the country

(b) *Quantitative random variables*

(*i*) Discrete quantitative random variables

- Number of people attending a conference
- Number of eggs in the refrigerator
- Number of children at a summer camp

(*ii*) Continuous quantitative random variables

- Heights of models in a beauty contest
- Weights of people joining a diet programme
- Lengths of steel bars produced in a given production run

Check Your Progress

1. What is sample process?
2. Define the term random sampling.
3. Elaborate on the stratified random sampling.
4. Explain the variables.
5. What is random variable?
6. Give the examples of qualitative random variables.

6.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population. It should be clear to the researcher that a sample is studied not for its own sake, but the basic objective of its study is to draw inference about the population.
2. Random sampling is refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term 'Arbitrary' for 'Random'.

3. In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum. If the ratio of items between various strata in the population matches with the ratio of corresponding items between various strata in the sample, it is called proportionate stratified sampling; otherwise, it is known as disproportionate stratified sampling.
4. A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.
5. A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response.
6. Qualitative Random Variables
    - Sex of students in the class
    - Political affiliation of a faculty member in the college
    - Opinions of economists regarding the economic conditions in the country

6.7 SUMMARY

- Under census or complete enumeration survey method, data is collected for each and every unit (e.g., person, consumer, employee, household, organization) of the population or universe which are the complete set of entities and which are of interest in any particular situation.
- Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population.
- The process of sampling involves selection of a sample based on a set of rules, collection of information and making an inference about the population. It should be clear to the researcher that a sample is studied not for its own sake, but the basic objective of its study is to draw inference about the population.
- Sampling is a tool which helps us know the characteristics of the universe or the population by examining only a small part of it.
- The values obtained from the study of a sample, such as the average and dispersion are known as ‘Statistics’ and the corresponding such values for the population are called ‘Parameters’.

NOTES

- The law of statistical regularity states that a moderately large number of items chosen at random from a large group are almost sure on the average to possess the characteristics of the large group. By random selection, we mean a selection where each and every item of the population has an equal chance of being selected.
- The law of inertia of large numbers states that, other things being equal, larger the size of the sample, more accurate the results are likely to be.
- A sound sampling procedure should result in a representative, adequate and homogeneous sample while ensuring that the selection of items should occur independently of one another.
- Random sampling is refers to that sampling technique in which each and every unit of the population has an equal chance of being selected in the sample. One should not mistake the term 'Arbitrary' for 'Random'.
- A random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample. Furthermore, it is easy to assess the accuracy of the sampling estimates because sampling errors follow the principles of chance.
- The sample size requirements would be usually larger under random sampling than under stratified random sampling, to ensure statistical reliability. It may escalate the cost of collecting data as the cases selected by random sampling tend to be too widely dispersed geographically.
- Stratified random sampling method, the universe to be sampled is subdivided (stratified) into groups which are mutually exclusive but collectively exhaustive based on a variable known to be correlated with the variable of interest.
- Stratified random sampling method differs from simple random sampling in that in the latter the sample items are chosen at random from the entire universe. In stratified random sampling, the sampling is designed in such a way that a designated number of items is chosen from each stratum.
- A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.
- There are two types of variables. One type is known as discrete variable and the other as continuous variable. A discrete variable takes whole number values and consists of distinct, recognizable individual elements that can be counted, such as the number of books in a library.
- A continuous variable is a variable whose values can theoretically take on an infinite number of values within a given range of values.
- A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response.

- A random variable may be qualitative or quantitative in nature. The qualitative random variables yield categorical responses so that the responses fit into one category or another.

NOTES

6.8 KEY WORDS

- **Population:** In statistical terms, a population is the totality of things under consideration.
- **Sampling:** Sampling is simply a process of learning about the population on the basis of a sample drawn from it. Thus, in any sampling technique, instead of every unit of the universe, only a part of the universe is studied and the conclusions are drawn on that basis for the entire population.
- **Random method:** A random method, it is independent of personal bias creeping into the analysis besides enhancing the representativeness of the sample.
- **Variable:** A variable is any characteristic which can assume different values. Age, height, IQ, and so on are all variables since their values can change when applied to different people.
- **Random variable:** A random variable is a phenomenon of interest in which the observed outcomes of an activity are entirely by chance, are absolutely unpredictable and may differ from response to response.

6.9 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the sampling.
2. Define the term random sampling.
3. Elaborate on the stratified random sampling.
4. What is variable?
5. Define the random variable.
6. Distinguish between the qualitative and quantitative random variables.

Long-Answer Questions

1. Discuss briefly about the various types of sampling with the help of examples.
2. Analyse the random sampling.
3. Describe the stratified random sampling.
4. Explain the types of variables. Give appropriate examples.
5. Differentiate between the continuous and discontinuous variables.

6.10 FURTHER READINGS

NOTES

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 7 PRESENTATION OF DATA

NOTES

Structure

7.0 INTRODUCTION

Presentation of data involves the use of a variety of different graphical techniques to visually show the reader the relationship between different data sets, to emphasise the nature of a particular aspect of the data or to geographically 'Place' data appropriately on a map.

A Single-Line Diagram (SLD), also sometimes called one-line diagram, is a simplified notation for representing a three-phase power system. The one-line diagram has its largest application in power flow studies. Bar graphs are the pictorial representation of data (generally grouped), in the form of vertical or horizontal rectangular bars, where the length of bars are proportional to the measure of data. They are also known as bar charts. Bar graphs are one of the means of data handling in statistics.

A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

Frequency polygons are a graphical device for understanding the shapes of distributions. They serve the same purpose as histograms, but are especially helpful for comparing sets of data. Frequency polygons are also a good choice for displaying cumulative frequency distributions. A pie diagram (or a circle chart) is a circular statistical graphic, which is divided into slices to illustrate numerical proportion.

In this unit, you will study about the line and bar diagram, histogram, polygon, pie diagram.

NOTES

7.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the line and bar diagram
- Explain the histogram
- Interpret the polygon
- Illustrate the pie diagram

7.2 LINE AND BAR DIAGRAM

Line Diagram

Here the points are plotted on paper (or graph paper) and joined by straight lines. Generally, continuous variables are plotted by the line graph.

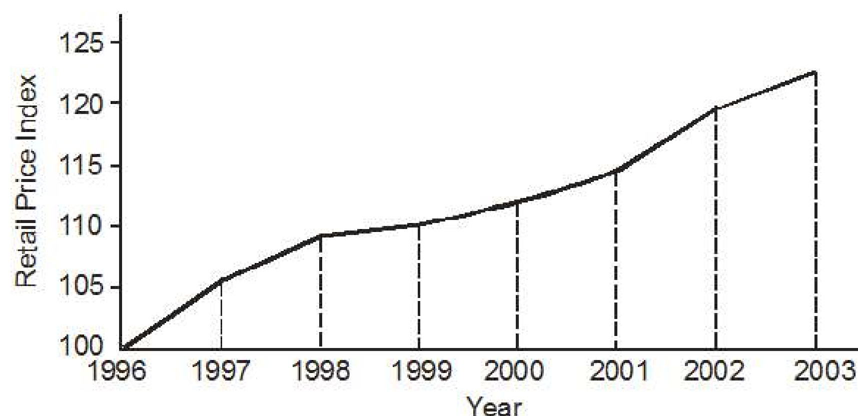
Example 7.1: The monthly averages of Retail Price Index from 1996 to 2003 (Jan. 1996 = 100) were as follows:

Year	1996	1997	1998	1999	2000	2001	2002	2003
Retail Price Index	100	105.8	109.0	109.6	110.7	114.5	119.3	122.3

Draw a diagram to display these figures.

Solution: Here years are plotted along the horizontal line and the retail price index along the vertical line.

Erect perpendiculars to horizontal line from the points marked as retail price index for the years 1997, 1998,…. 2003 and cut off these ordinates according to the given data and thus various points will be plotted on the paper. Join these points by straight lines.



NOTES

The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data. These important features are more simply presented in the form of graphs. Also, diagrams facilitate comparisons between two or more sets of data.

The diagrams should be clear and easy to read and understand. Too much information should not be shown in the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self-explanatory title dealing with the subject matter. The scale of the presentation should be chosen in such a way that the resulting diagram is of appropriate size. The intervals on the vertical as well as the horizontal axis should be of equal size; otherwise, distortions would occur.

Diagrams are more suitable to illustrate the data which is discrete, while continuous data is better represented by graphs. The following are the diagrammatic and graphic representation methods that are commonly used.

Diagrammatic Representation

(a) Bar diagram; (b) Pie chart; (c) Pictogram

Bar Diagrams

One Dimensional Bar Diagrams

Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.

Example 7.2: Suppose that the following were the gross revenues (in \$100,000.00) for a company *XYZ* for the years 1989, 1990 and 1991.

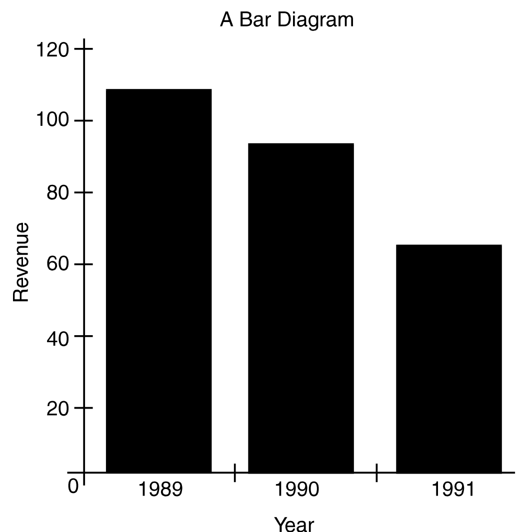
Year	Revenue
1989	110
1990	95
1991	65

Construct a bar diagram for this data.

NOTES

Solution:

The bar diagram for this data can be constructed as follows with the revenues represented on the vertical axis and the years represented on the horizontal axis.



Two Dimensional Bar Diagrams

When each figure is made up of two or more component figures the bars may be sub-divided into components. Too many components should not be shown.

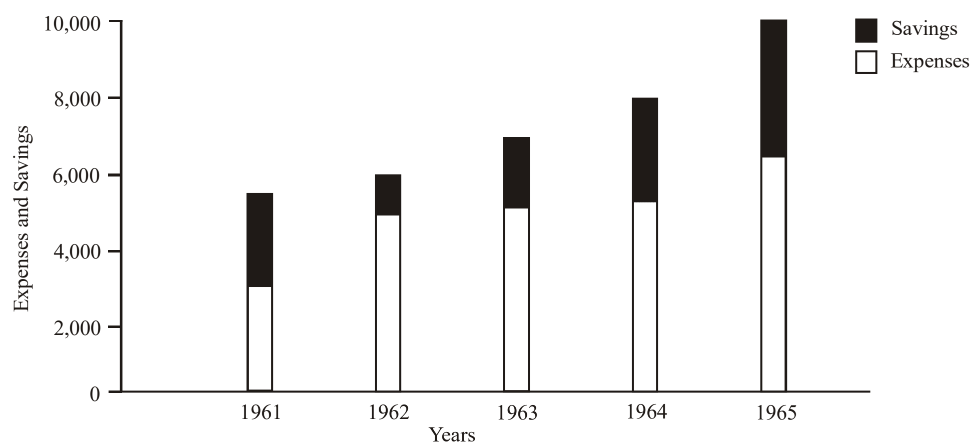


Fig. 7.1 *Component Bar Chart Showing Expenses and Savings of Mr. X*

Table 7.1 Showing Annual Income, Expenses and Savings of Mr. X

Year	Amounts in Rs of			Percentages of		
	Income	Expenses	Savings	Income	Expenses	Savings
1961	5000	3000	2000	100.0	60.0	40.0
1962	6000	5000	1000	100.0	83.3	16.7
1963	7000	5000	2000	100.0	71.4	28.6
1964	8000	5000	3000	100.0	62.5	37.5
1965	10000	6000	4000	100.0	60.0	40.0

NOTES

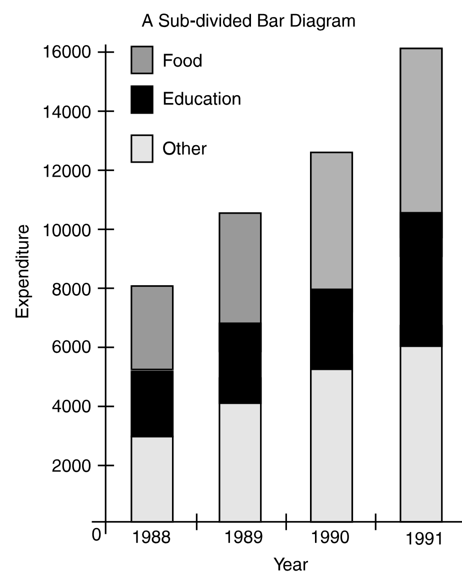
The bars drawn can be further subdivided into components depending upon the type of information to be shown in the diagram. This will be clear by the following example in which we are presenting three components in a bar.

Example 7.3: Construct a subdivided bar chart for the three types of expenditures in dollars for a family of four for the years 1988, 1989, 1990 and 1991 as given as fllows:

Year	Food	Education	Other	Total
1988	3000	2000	3000	8000
1989	3500	3000	4000	10500
1990	4000	3500	5000	12500
1991	5000	5000	6000	16000

Solution:

The subdivided bar chart would be as follows:



NOTES

Percentage Component Bars or Divided Bar Charts

When in the above case the component lengths represent the percentages (instead of the actual amounts) of each component we get percentage component bar charts. The heights of all the bars will be the same.

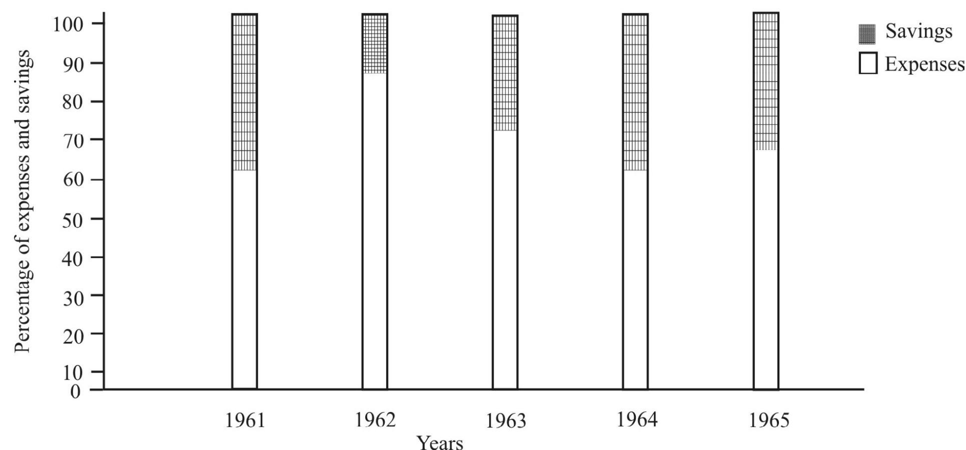


Fig. 7.2 *Percentage Component Bar Chart Showing Expenses and Savings of Mr. X*

Multiple Bar Charts

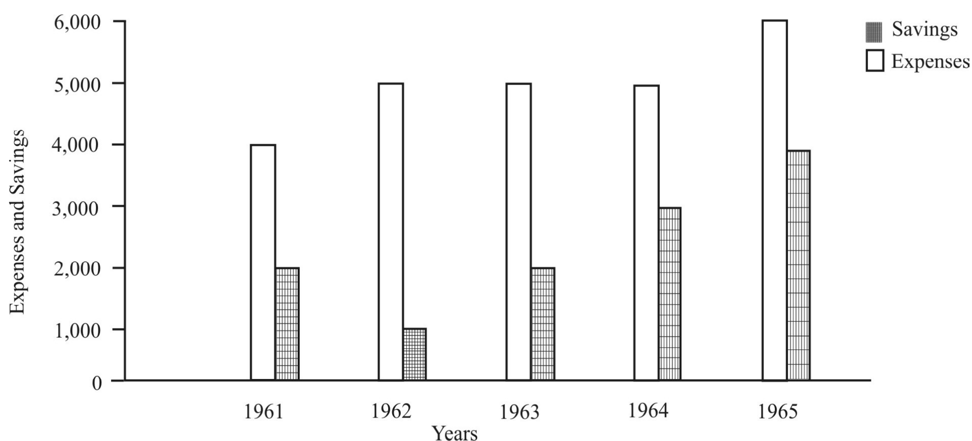


Fig. 7.3 *Multiple Bar Chart Showing Expenses and Savings of Mr. X*

Here the interrelated component parts are shown in adjoining bars, coloured or marked differently, thus allowing comparison between different parts.

These charts can be used if the overall total is not required. Some charts given earlier show totals also.

Pictogram

Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.

News magazines are very fond of presenting data in this form. For example, in comparing the strength of the armed forces of USA and Russia, they will simply show sketches of soldiers where each sketch may represent 100,000 soldiers. Similar comparison for missiles and tanks is also done.

NOTES

7.3 HISTOGRAM

A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.

The word histogram is derived from the Greek word histos which means 'anything set upright' and 'gramma' which means 'drawing, record, writing'. It is considered as the most important basic tool of statistical quality control process.

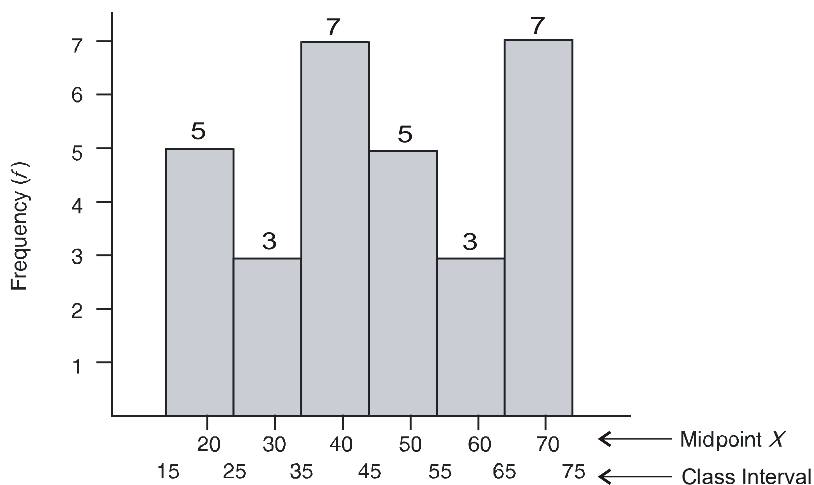
In this type of representation the given data are plotted in the form of a series of rectangles. Class intervals are marked along the X-axis and the frequencies along the Y-axis according to a suitable scale. Unlike the bar chart, which is one-dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important. A histogram is constructed from a frequency distribution of a grouped data where the height of the rectangle is proportional to the respective frequency and the width represents the class interval. Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.

As an example, let us construct a histogram for our example of ages of 30 workers. For convenience sake, we will present the frequency distribution along with the mid-point of each interval, where the mid-point is simply the average of the values of the lower and upper boundary of each class interval. The frequency distribution table is shown as follows:

Class Interval (years)	Mid-point	(f)
15 and upto 25	20	5
25 and upto 35	30	3
35 and upto 45	40	7
45 and upto 55	50	5
55 and upto 65	60	3
65 and upto 75	70	7

The histogram of this data would be shown as follows:

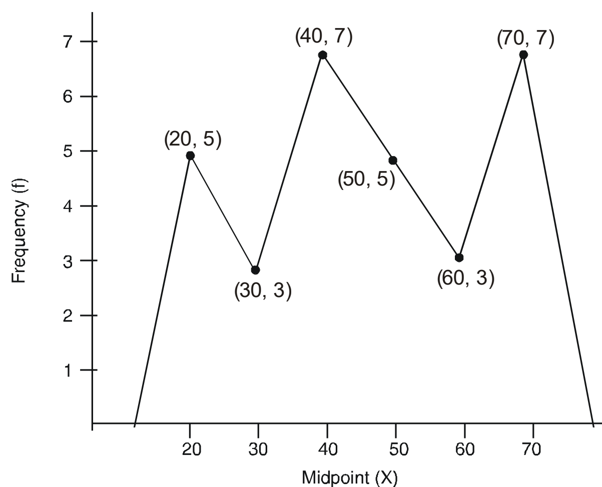
NOTES



7.4 POLYGON

A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies and these plotted points are joined together by straight lines. Since the frequencies generally do not start at zero or end at zero, this diagram as such would not touch the horizontal axis. However, since the area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero, so that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.

We can construct the frequency polygon from the table presented for the ages of 30 workers as follows:



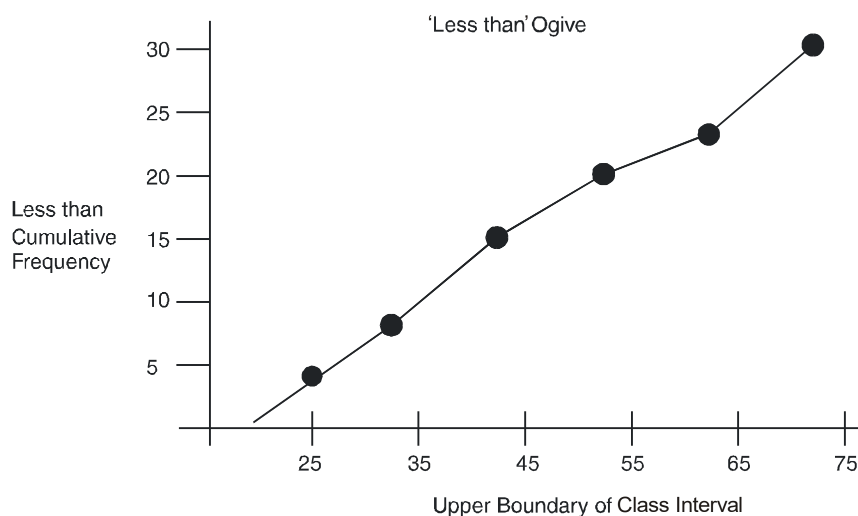
NOTES

Cumulative Frequency Curve (Ogives)

The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.

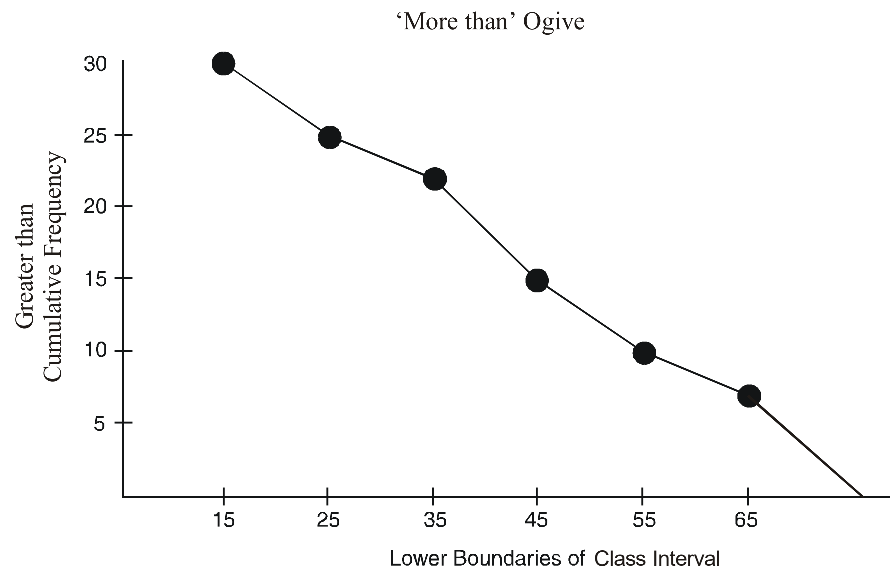
Class Interval (years)	Mid-point	(*f*)	Cum. Freq. (less than)	Cum. Freq. (greater than)
15 and upto 25	20	5	5 (less than 25)	30 (more than 15)
25 and upto 35	30	3	8 (less than 35)	25 (more than 25)
35 and upto 45	40	7	15 (less than 45)	22 (more than 35)
45 and upto 55	50	5	20 (less than 55)	15 (more than 45)
55 and upto 65	60	3	23 (less than 65)	10 (more than 55)
65 and upto 75	70	7	30 (less than 75)	7 (more than 65)

(*i*) Less than ogive. In this case less than cumulative frequencies are plotted against upper boundaries of their respective class intervals.



(*ii*) Greater than ogive. In this case greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals.

NOTES



These ogives can be used for comparison purposes. Several ogives can be drawn on the same grid, preferably with different colours for easier visualization and differentiation.

Although, diagrams and graphs are a powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.

7.5 PIE DIAGRAM

This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

Example 7.4: The following figures relate to the cost of the construction of a house. The various components of cost that go into it are represented as percentages of the total cost.

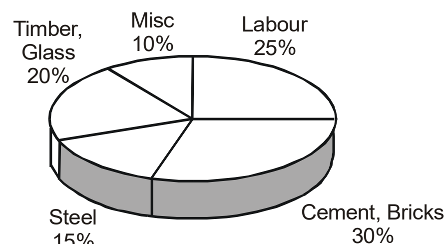
Item	*% Expenditure*
Labour	25
Cement, bricks	30
Steel	15
Timber, glass	20
Miscellaneous	10

Construct a pie chart for the above data.

NOTES

Solution:

The pie chart for this data is presented as follows:



Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

7.6 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.
2. Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
3. A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
4. The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive. Both these ogives are constructed based upon the following table of our example of 30 workers.
5. The type of pie diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it. The size of the slice represents the proportion of the component out of the whole.

7.7 SUMMARY

- The data we collect can often be more easily understood for interpretation if it is presented graphically or pictorially. Diagrams and graphs give visual indications of magnitudes, groupings, trends and patterns in the data.

NOTES

- The diagrams should be clear and easy to read and understand. Too much information should not be shown in the same diagram; otherwise, it may become cumbersome and confusing. Each diagram should include a brief and self-explanatory title dealing with the subject matter.
- Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram.
- Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value.
- A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
- The word histogram is derived from the Greek word histos which means 'Anything Set Upright' and 'Gramma' which means 'Drawing, Record, and Writing'. It is considered as the most important basic tool of statistical quality control process.
- The type of histogram representation the given data are plotted in the form of a series of rectangles. Class intervals are marked along the X-axis and the frequencies along the Y-axis according to a suitable scale.
- the bar chart, which is one dimensional, meaning that only the length of the bar is important and not the width, a histogram is two-dimensional in which both the length and the width are important.
- A histogram is constructed from a frequency distribution of a grouped data where the height of the rectangle is proportional to the respective frequency and the width represents the class interval.
- Each rectangle is joined with the other and any blank spaces between the rectangles would mean that the category is empty and there are no values in that class interval.
- A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies or these plotted points are joined together by straight lines.
- The area under the entire curve is the same as that of a histogram which is 100 per cent of the data presented, the curve can be enclosed so that the starting point is joined with a fictitious preceding point whose value is zero, so that the start of the curve is at horizontal axis and the last point is joined with a fictitious succeeding point whose value is also zero, so that the curve ends at the horizontal axis. This enclosed diagram is known as the frequency polygon.
- The cumulative frequency curve or ogive is the graphic representation of a cumulative frequency distribution. Ogives are of two types. One of these is less than and the other one is greater than ogive.

- Less than ogive. In this case less than cumulative frequencies are plotted against upper boundaries of their respective class intervals.
- Greater than ogive. In this case greater than cumulative frequencies are plotted against the lower boundaries of their respective class intervals.
- Diagrams and graphs are a powerful and effective media for presenting statistical data, they can only represent a limited amount of information and they are not of much help when intensive analysis of data is required.
- This type of diagram enables us to show the partitioning of a total into its component parts. The diagram is in the form of a circle and is also called a pie because the entire diagram looks like a pie and the components resemble slices cut from it.
- Pie charts are very useful for comparison purposes, especially when there are only a few components. If there are too many components, it may become confusing to differentiate the relative values in the pie.

7.8 KEY WORDS

- **Bar diagrams:** Bars are simply vertical lines where the lengths of the bars are proportional to their corresponding numerical values. The width of the bar is unimportant but all bars should have the same width so as not to confuse the reader of the diagram. Additionally, the bars should be equally spaced.
- **Pictogram:** Pictogram means presentation of data in the form of pictures. It is quite a popular method used by governments and other organizations for informational exhibitions. Its main advantage is its attractive value. Pictograms stimulate interest in the information being presented.
- **Histogram:** A histogram is the graphical description of data and is constructed from a frequency table. It displays the distribution method of a data set and is used for statistical as well as mathematical calculations.
- **Polygon:** A frequency polygon is a line chart of frequency distribution in which either the values of discrete variables or mid-points of class intervals are plotted against the frequencies or these plotted points are joined together by straight lines.

7.9 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is line diagram?
2. Illustrate the bar diagram.

NOTES

3. Define the term pictogram.
4. Elaborate on the histogram.
5. What do you understand by polygon?
6. What is cumulative frequency curve?
7. Illustrate the pie diagram.

Long-Answer Questions

1. Discuss briefly about the line and bar diagram.
2. Briefly explain about the histogram with the help of examples.
3. Analyse the polygon. Give the appropriate examples.
4. Describe the pie diagram.

7.10 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

NOTES

BLOCK - III
MEASURES OF CENTRAL TENDENCY AND MEASURE OF DISPERSION

UNIT 8 MEAN, MEDIAN AND MODE

Structure

8.0 INTRODUCTION

In probability and statistics, the population mean, or expected value, is a measure of the central tendency either of a probability distribution or of a random variable characterized by that distribution. In statistics and probability theory, the median is the value separating the higher half from the lower half of a data sample, a population, or a probability distribution. For a data set, it may be thought of as the 'Middle' value. The basic feature of the median in describing data compared to the mean (often simply described as the 'Average') is that it is not skewed by a small proportion of extremely large or small values, and therefore provides a better representation of a 'Typical' value. The mode is the value that appears most often in a set of data values. If X is a discrete random variable, the mode is the value x (i.e., $X = x$) at which the probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

In statistics, dispersion (also called variability, scatter, or spread) is the extent to which a distribution is stretched or squeezed. In statistics, the means of range 'A Set of Data' is the difference between the largest and smallest values. It can give you a rough idea of how the outcome of the data set will be before you look at it actually Difference here is specific, the range of a set of data is the result of subtracting the smallest value from largest value. In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. In other words, it measures how far a set of numbers is spread out from

their average value. Variance has a central role in statistics, where some ideas that use it include descriptive statistics, statistical inference, hypothesis testing, and goodness of fit, and Monte Carlo sampling. Variance is an important tool in the sciences, where statistical analysis of data is common.

NOTES

In this unit, you will study about the mean, median and mode: dispersion, range, variance of SD, SE, and CV.

8.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the dispersion, range, variance of SD, SE, and CV

8.2 DISPERSION: RANGE, VARIANCE, SD, SE AND CV

Biostatistics refers to the calculation and analysis of the data obtained in biological studies, researches or experiments. Summarizing the biological data obtained after a measurement variable requires a number for representing the 'middle' of a set of numbers, termed as the 'Statistic of Central Tendency' or 'Statistic of Location', along with a measure of the 'Spread' of the numbers. In statistics, a central tendency or measure of central tendency is a central or typical value for a probability distribution. It is also termed as a center or location of the distribution. Occasionally, the measures of central tendency are often termed as averages. The most common measures of central tendency are the arithmetic mean, the median and the mode. A central tendency can be calculated for either a finite set of values or for a theoretical distribution, such as the normal distribution.

Mean or Arithmetic Mean

Mean or arithmetic mean, median, and mode are the different measures of centre in a numerical data set. It summarizes a dataset with a single number to represent a 'typical/unique' data point from the dataset.

The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'the mean' or 'the average' of the dataset. Basically, it is the 'average' number obtained by adding all data points and then dividing by the number of data points.

For example, the mean of 4, 1 and 7 is $(4 + 1 + 7) / 3 = 12 / 3 = 4$.

Mean is, thus, the most commonly used form of all the averages. It is the value which is obtained by dividing the aggregate of various items of the same series by the total number of observations.

Therefore, the mean or arithmetic mean is the sum of all of the data points divided by the number of data points.

Mean = Sum of Data / # of Data Points

The following is the formula:

$$\text{Mean} = \frac{\sum x_i}{n}$$

Example 8.1. Find the mean of the following given data:

1, 3, 4, 5, 7, 9

Solution: We first add the data as follows:

1 + 2 + 3 + 4 + 5 + 7 + 8 = 30

There are 6 data points, hence as per the formula = 30 / 6 = 5

Therefore the mean is 5.

Calculation of Mean for Ungrouped Data

When observations are denoted by x values showing $x_1, x_2, x_3 \ldots x_n$, then the total number of observations is calculated by summing up the observations and dividing the sum by the total number of observations (n).

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}$$

Example 8.2: Find out the mean or average pod length of the plant from the following data.

The pod length of the ten pods of a plant are as follows:

5.2 cm 5.3 cm 5.6 cm 5.7 cm 5.4 cm

5.2 cm 5.3 cm 5.3 cm 5.4 cm 5.2 cm

Solution: The mean or average pod length of the plant is calculated as,

$$\bar{x} = \frac{\sum x}{n} = \frac{5.2 + 5.3 + 5.6 + 5.7 + 5.4 + 5.2 + 5.3 + 5.3 + 5.4 + 5.2}{10} cm$$

$$= \frac{53.6}{10} cm = 5.36 cm$$

Calculation of Mean for Grouped Data

When the series is discrete, each value of the variable is multiplied by their respec-tive frequencies, and the sum of all values is divided by total number of frequencies. Variable x has the values like $x_1, x_2, x_3, \ldots, x_n$ and their frequencies are f_1 , f_2, f_3, …, f_n, respectively.

The mean or arithmetic mean is calculated using the formula:

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + f_3x_3 + \cdots + f_nx_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum fx}{\sum f}$$

NOTES

When the series is continuous, the arithmetic mean is calculated after taking the midpoint value of class intervals.

$$\bar{x} = \frac{\sum fm}{\sum f}$$

Where,

$\bar{x}$ = Arithmetic Mean

$\sum$fm = Sum Values of Midpoint Value Multiplied by their Frequencies

$\sum$f = Sum of Frequencies

m = Mid Points of Various Class Intervals

Example 8.3: An observation on 32 Balsam plants shows the following data.

Calculate the arithmetic mean.

No. of flowers/plant (x)	4	5	6	7	8	9
No. of plants (f)	3	5	6	9	5	4

Solution: We calculate the arithmetic mean as follows:

No. of flowers/plant (x)	No. of plants (f)	fx
4	3	12
5	5	25
6	6	36
7	9	63
8	5	40
9	4	6
	Σf = 32	Σfx = 212

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{212}{32} = 6.62(\text{approx})$$

The average number of flowers/plant is 6.62.

The average mean or arithmetic mean is calculated as follows:

No. of pods/plant	Mid points of class (m)	No. of plants frequency (f)	m.f.
15-17	16	5	80
18-20	19	6	114
21-23	22	8	176
24-26	25	12	300

27-29	28	22	616
30-32	31	18	558
33-35	34	15	510
36-38	37	9	333
39-41	40	5	200
		Σf = 100	Σmf = 2.887

NOTES

$$\text{Arithmetic Mean} = \overline{x} = \frac{\sum mf}{\sum f} = \frac{2.887}{100} = 28.87.$$

The arithmetic mean is 28.87.

Merits, Demerits and Uses of Arithmetic Mean

Merits

1. The formula to calculate the arithmetic mean is very simple and it is easily understood.
2. The arithmetic mean is firmly defined mathematical formula hence the same result may come on repeated calculations.
3. The calculation of arithmetic mean is based on all the observations.
4. The arithmetic mean is least affected by sampling fluctuation.
5. The arithmetic mean balances the value on either side.
6. The arithmetic mean is the best measure to compare two or more series.
7. Arithmetic mean is totally dependent on values and not on the position.

Demerits

1. The arithmetic mean cannot be calculated if all the values are not known.
2. In arithmetic mean the extreme values have greater effect on mean.
3. The qualitative data cannot be measured using this method.

Uses

1. The arithmetic mean is mostly used in practical statistics.
2. Mean helps to calculate many other estimates in statistics.
3. The arithmetic mean is most popular method of any measurement used by common people to get the average of any data.

Median

The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then

taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.

NOTES

For example, the median of 4, 1, and 7 is 4 because when the numbers are arranged in order then we have the sequence (1, 4, 7) in which 4 is the middle number.

The median of a distribution is defined as the value of that variable which divides the total frequency into two equal parts when the series is arranged in ascending or descending order of magnitude. So in a distribution, half of the values remain below median value and half of the values remain above the median value.

Finding the Median

- Arrange the data points from smallest to largest.
- If the number of data points is odd, the median is the middle data point in the list.
- If the number of data points is even, the median is the average of the two middle data points in the list.

Example 8.4: Find the median of the following data:

1, 4, 2, 5, 0

Solution: First arrange the data in order form as follows:

0, 1, 2, 4, 5

Because there is an odd number of data points, hence the median is the middle data point, i.e., 2.

0, 1, **2**, 4, 5

Therefore the median is 2.

Example 8.5. Find the median of the following data:

10, 40, 20, 50

Solution: First arrange the data in order form as follows:

10, 20, 40, 50

In this case, there is an even number of data points, hence the median is the average of the middle two data points, i.e., 20 and 40.

10, **20**, **40**, 50

Therefore,

Median = 20 + 40 / 2 = 60 / 2 = 30

The median in this case is 30.

NOTES

Median Value for Ungrouped Data

Median value is the value of the n+1 / 2 item. But this formula is applicable only when item number is odd. But when the item number is even, then the median value is calculated by the mean value of n/2th and (n / 2+1)th items,

$$\therefore \text{ Median} = \frac{\frac{n}{2}\text{th value} + \left(\frac{n}{2}+1\right)\text{th value}}{2}$$

Example 8.6: Calculate the median number of flowers in the following observation obtained from garden plants.

Plant no.	1	2	3	4	5	6	7	8
No. of flowers	20	17	25	18	23	21	16	26

Solution: The median is calculated as follows:

Item no.	No. of flowers/plant Ascending	No. of flowers/plant Descending
1	16	26*
2	17	25
3	18	23
4	20	21
5	21	20
6	23	18
7	25	17
8	26*	16

The observations are arranged in both ascending and descending order. In case of observation of 7 plants the * marked item no. should not be considered.

If we take 7 observations, then the median value will be value of 7+1/2th, i.e., 4th item, i.e., 20.

If we take 8 observations, then the median value will be the mean of 8/2th and 8/2 +1th item, i.e., 21.

Mean of 4th and 5th item, i.e., mean of 20 and 21 which is 20.5.

Median Value for Grouped Data

For grouped data, the classes are arranged according to the ascending order and respective frequencies are written against them. The frequencies are then cumulated and position of the median is calculated by the same formula. The median value is the mid value of the class in which the median item value is placed. Consider the following table showing the class interval, mid value, frequency and cumulative frequency for number of pods.

NOTES

Class interval	Mid value	Frequency	Cumulative frequency
15-17	16	5	5
18-20	19	6	11
21-23	22	8	19
24-26	25	12	31
27-29	28	22	53
30-32	31	18	71
33-35	34	15	86
36-38	37	9	95
39-41	40	5	100

Because the total number of variables is 100, hence the median value will be the value which is in between the value of 50th and 51st item value.

50th and 51st item value is in the class interval 27-29 (No. of Pods).

Therefore, the median value is 28 of this observation.

Merits and Demerits of Median

Merits

In normal distribution, the median value is near the mean value which is easier to calculate. This value eliminates the effect of extreme items, since they are not taken into account for its calculation, hence only the middle items must be known.

Demerits

When the distribution is irregular then the median value is not considered as the true representative of the series. In case of grouped data also, the precision is lost, hence this value is not significant for further analysis.

Mode

The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.

Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.

For example, the mode of the dataset {4, 2, 4, 3, 2, 2} is 2 because it is occurring three times, which is more than any other number.

Example 8.7: Find the mode of the following data:

0, 0, 1, 1, 1, 1, 1, 1, 2, 2, 2, 3, 5

The most occurring value in the dataset is,

0, 0, **1, 1, 1, 1, 1, 1**, 2, 2, 2, 3, 5

Therefore, the mode is 1.

In a frequency distribution, 'mode' is defined as, 'The value of the variable for which the frequency is maximum'. From the definition it is clear that mode

cannot be determined from a series of individual observation, and it always depends on the frequency of occurrence of any item. When the concentration of data gives only one peak then the distribution is unimodal, but if the data concentrates at two or more points on a scale of values, then the series is called bimodal or multimodal.

NOTES

In the Example 8.3, we obtain the maximum frequency in case of variable value 7. Therefore the mode value of this observation is 7. This type of distribution is called unimodal distribution. In Example 3, the maximum frequency (22) is observed in case of class value 27-29, hence the mid value of this class is 28. Consequently, the mode value of this observation is 28.

Example 8.8: The observation on 30 Balsam plants shows the following data. Calculate the mode from this observation.

No. of flowers/plant (x)	3	4	5	6	7	8	9	10
No. of plants (f)	1	3	2	8	5	8	2	1

Solution: Here the mode value cannot be calculated by just assessment, as the maximum frequency is observed in case of two values of variable 6 and 8. Therefore to determine the modal class, the data is grouped as follows.

If 2 values are taken together then the grouped data can be arranged in the following manner:

Class value	Mid value (m)	frequency
3-4	3.5	4
5-6	5.5	10
7-8	7.5	13
9-10	9.5	3

Here the modal class is 7-8, where mid value is 7.5, so the mode value of this distribution is 7.5. This type of distribution is called bimodal distribution.

Merits and Demerits of Mode

Merits

- The mode value avoids the effects of extreme items. The value is obtained by mere assessment of data.
- All values may not to be known, it refers to a measurement which is most usual and most likely variate.
- The bimodal or multimodal distribution gives good indication of the heterogeneity of the population.

Demerits

This value does not require any kind of calculation. It becomes difficult sometimes to mention the bimodal or multimodal distribution. This value is less dependable as all observations in a series do not have any influence on the value.

8.2.1 Standard Deviation (SD) and Standard Error (SE) of Mean

NOTES

The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. Basically, it is calculated as the square root of variance by determining the variation between each data point relative to the mean. If the data points are further from the mean, there is a higher deviation within the dataset; thus, the more spread out the data, the higher the standard deviation.

Standard deviation is the most commonly used measure of dispersion of data around a mean - described more frequently than the variance. Arithmetically, standard deviation is defined as the square root of the variance.

The Standard Deviation (SD) is a measure of how spread out numbers are.

The symbol for Standard Deviation is σ (the Greek letter sigma) and the formula for Standard Deviation is:

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Say we have a bunch of flowers represented using numbers as 9, 2, 5, 4, 12, 7, 8, and 11.

To calculate the Standard Deviation of these numbers:

1. Calculate the Mean (the simple average of the numbers).
2. Then for each number, subtract the Mean and Square the result.
3. Then calculate the Mean of those Squared Differences.
4. Take the square root to obtain the result.

Example 8.9: Assume that there are 20 rose bushes and the number of flowers on each bush is,

9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4

Calculate the Standard Deviation.

Solution: The Standard Deviation is calculated as follows.

Step 1. Calculate the mean.

In the formula above m is the mean of all the values.

For example, we have the dataset 9, 2, 5, 4, 12, 7, 8, 11, 9, 3, 7, 4, 12, 5, 4, 10, 9, 6, 9, 4.

The mean for this is:

9 + 2 + 5 + 4 + 12 + 7 + 8 + 11 + 9 + 3 + 7 + 4 + 12 + 5 + 4 + 10 + 9 + 6 + 9 + 4 / 20

= 140 / 20 = 7

Therefore, $\mu = 7$

Step 2. Then for each number: Subtract the Mean and Square the result.

NOTES

As per the formula,

$(x_i - \mu)2$

Here x_i refers to the individual values of x which are 9, 2, 5, 4, 12, 7, etc.

In other words $x_1 = 9$, $x_2 = 2$, $x_3 = 5$, and so on.

For each value, subtract the mean and square the result as follows:

$(9 - 7)^2 = (2)^2 =$ **4**

$(2 - 7)^2 = (-5)^2 =$ **25**

$(5 - 7)^2 = (-2)^2 =$ **4**

$(4 - 7)^2 = (-3)^2 =$ **9**

$(12 - 7)^2 = (5)^2 =$ **25**

$(7 - 7)^2 = (0)^2 =$ **0**

$(8 - 7)^2 = (1)^2 =$ **1**

And so on.

The final result is:

4, 25, 4, 9, 25, 0, 1, 16, 4, 16, 0, 9, 25, 4, 9, 9, 4, 1, 4, 9

Step 3. The.n calculate the mean of those squared differences.

To calculate the mean, add up all the values then divide by how many.

First add up all the values from the previous step using 'Sigma' - Σ.

To add up all the values from 1 to N, where $N = 20$ in this case because there are 20 values:

$$\sum_{i=1}^{N} (x_i - \mu)^2$$

Which means that sum all values from $(x_1 - 7)^2$ to $(x_N - 7)^2$

We have already calculated $(x_1 - 7)^2 = 4$ in the previous step. Therefore, to sum them up:

= 4 + 25 + 4 + 9 + 25 + 0 + 1 + 16 + 4 + 16 + 0 + 9 + 25 + 4 + 9 + 9 + 4 + 1 + 4 + 9

= **178**

But this is not yet the mean. To obtain the mean, we have multiply by $1/N$ (the same as dividing by N):

$$\frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)^2$$

Mean of Squared Differences = (1/20) × 178 = 8.9

This value is called the **'Variance'**.

Step 4. Taking the square root:

Therefore the **Standard Deviation** 'σ' = √ (8.9) = 2.983.

Standard Error of the Mean (SEM)

The Standard Deviation (SD) measures the amount of variability, or dispersion, for a subject set of data from the mean, while the Standard Error of the Mean (SEM) measures how far the sample mean of the data is likely to be from the true mean. The SEM is always smaller than the SD. Both the 'Standard Deviation' and 'Standard Error (SE)' are often used in experimental studies. In these studies, the Standard Deviation (SD) and the estimated Standard Error of the Mean (SEM) are used to present the characteristics of sample data and to explain statistical analysis results. Alternatively, SD indicates how accurately the mean represents sample data. However, the meaning of SEM includes statistical inference based on the sampling distribution. SEM is the SD of the theoretical distribution of the sample means (the sampling distribution).

Calculating Standard Error of the Mean Standard Deviation 'σ' is:

$$\sqrt{\frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1}}$$

Variance = σ^2

$$\text{Standard Error}\left(\sigma_{\bar{x}}\right) = \frac{\sigma}{\sqrt{n}}$$

Where,

$\bar{x}$ = Sample Mean

n = Sample Size

Therefore, the Standard Error of the Mean (SEM) is calculated by taking the standard deviation and dividing it by the square root of the sample size.

The formula for the SD includes the following steps:

Step 1: Take the square of the difference between each data point and the sample mean, finding the sum of those values.

Step 2: Divide that sum by the sample size minus one, which is the variance.

Step 3: Take the square root of the variance to obtain the SD.

Standard error validates the accuracy of a single sample or of the multiple samples by analysing deviation within the means. The SEM defines how precise the mean of the sample is vs. the true mean of the dataset. As the size of the sample data grows larger, the SEM decreases vs. the SD.

The standard error is defined as the measure of descriptive statistics. It represents the standard deviation of the mean within a dataset. Thus, it functions as a measure of variation for random variables, providing a measurement for the spread. The smaller the spread, the more accurate the dataset.

NOTES

8.2.2 Coefficient of Variation (CV)

Introduction

In probability theory and statistics, the **Coefficient of Variation (CV),** also known as **Relative Standard Deviation (RSD),** is a standardized measure of dispersion of a probability distribution or frequency distribution. It is often expressed as a percentage, and is defined as the ratio of the standard deviation σ to the mean μ (or its absolute value, $|\mu|$). The CV or RSD is widely used in analytical chemistry to express the precision and repeatability of an assay. It is also commonly used in fields, such as engineering or physics when doing quality assurance studies and ANalysis of VAriance (ANOVA) gauge Repeatability and Reproducibility (R&R). In addition, CV is utilized by economists and investors in economic models.

Definition

The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$. It shows the extent of variability in relation to the mean of the population. The coefficient of variation should be computed only for data measured on a **Ratio Scale**, that is, scales that have a meaningful zero and hence allow relative comparison of two measurements (i.e., division of one measurement by the other). The coefficient of variation may not have any meaning for data on an **Interval Scale.** For example, most temperature scales (e.g., Celsius, Fahrenheit, etc.) are interval scales with arbitrary zeros, so the computed coefficient of variation would be different depending on which scale you used. On the other hand, Kelvin temperature has a meaningful zero, the complete absence of thermal energy, and thus is a ratio scale. In plain language, it is meaningful to say that 20 Kelvin is twice as hot as 10 Kelvin, but only in this scale with a true absolute zero. While a Standard Deviation (SD) can be measured in Kelvin, Celsius, or Fahrenheit, the value computed is only applicable to that scale. Only the Kelvin scale can be used to compute a valid coefficient of variability.

Measurements that are log-normally distributed exhibit stationary CV; in contrast, SD varies depending upon the expected value of measurements.

A more robust possibility is the quartile coefficient of dispersion, half the interquartile range $(Q_3\text{-}Q_1)/2$ divided by the average of the quartiles, $(Q_1+Q_3)/2$.

In most cases, a CV is computed for a single independent variable (e.g., a single factory product) with numerous, repeated measures of a dependent variable (e.g., error in the production process). However, data that are linear or even logarithmically non-linear and include a continuous range for the independent variable with sparse measurements across each value (e.g., scatter-plot) may be amenable to single CV calculation using a maximum-likelihood estimation approach.

Example 8.10: A data set of [100, 100, 100] has constant values. Its standard deviation is 0 and average is 100, giving the coefficient of variation as

0 / 100 = **0**

NOTES

A data set of [90, 100, 110] has more variability. Its sample standard deviation is 10 and its average is 100, giving the coefficient of variation as

10 / 100 = **0.1**

A data set of [1, 5, 6, 8, 10, 40, 65, 88] has still more variability. Its standard deviation is 32.9 and its average is 27.9, giving a coefficient of variation of

32.9 / 27.9 = **1.18**

Check Your Progress

1. What is mean?
2. Give the demerits of arithmetic mean.
3. Explain the uses of arithmetic mean.
4. Define the term median.
5. Give the merits of median.
6. What do you understand by mode?
7. Elaborate on the Standard Deviation (SD).
8. Define the term Coefficient of Variation (CV).

8.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Mean or arithmetic mean, median, and mode are the different measures of centre in a numerical data set. It summarizes a dataset with a single number to represent a 'Typical/Unique' data point from the dataset.
2. Demerits of Arithmetic Mean
   - The arithmetic mean cannot be calculated if all the values are not known.
   - In arithmetic mean the extreme values have greater effect on mean.
   - The qualitative data cannot be measured using this method.
3. Uses of Arithmetic Mean
   - The arithmetic mean is mostly used in practical statistics.
   - Mean helps to calculate many other estimates in statistics.
   - The arithmetic mean is most popular method of any measurement used by common people to get the average of any data.

NOTES

4. The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.
5. In normal distribution, the median value is near the mean value which is easier to calculate. This value eliminates the effect of extreme items, since they are not taken into account for its calculation, hence only the middle items must be known.
6. The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times. Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.
7. The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
8. The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$. It shows the extent of variability in relation to the mean of the population.

8.4 SUMMARY

- Biostatistics refers to the calculation and analysis of the data obtained in biological studies, researches or experiments. Summarising the biological data obtained after a measurement variable requires a number for representing the 'middle' of a set of numbers, termed as the 'Statistic of Central Tendency' or 'Statistic of Location', along with a measure of the 'Spread' of the numbers. In statistics, a central tendency or measure of central tendency is a central or typical value for a probability distribution.
- Mean or arithmetic mean, median, and mode are the different measures of centre in a numerical data set. It summarizes a dataset with a single number to represent a 'Typical/Unique' data point from the dataset.
- The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'The Mean' or 'The Average' of the dataset. Basically, it is the 'Average' number obtained by adding all data points and then dividing by the number of data points.
- Mean is, thus, the most commonly used form of all the averages. It is the value which is obtained by dividing the aggregate of various items of the same series by the total number of observations.

NOTES

- When the series is discrete, each value of the variable is multiplied by their respective frequencies, and the sum of all values is divided by total number of frequencies.
- The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers. Therefore, the median is the middle point in a dataset, i.e., half of the data points are smaller than the median and half of the data points are larger.
- The median of a distribution is defined as the value of that variable which divides the total frequency into two equal parts when the series is arranged in ascending or descending order of magnitude. So in a distribution, half of the values remain below median value and half of the values remain above the median value.
- Median value for grouped data, the classes are arranged according to the ascending order and respective frequencies are written against them. The frequencies are then cumulated and position of the median is calculated by the same formula. The median value is the mid value of the class in which the median item value is placed.
- When the distribution is irregular then the median value is not considered as the true representative of the series. In case of grouped data also, the precision is lost, hence this value is not significant for further analysis.
- The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times.
- The mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset. There can be no mode, one mode, or multiple modes in a dataset.
- In a frequency distribution, 'Mode' is defined as, 'The value of the variable for which the frequency is maximum'. From the definition it is clear that mode cannot be determined from a series of individual observation, and it always depends on the frequency of occurrence of any item.
- The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance. Basically, it is calculated as the square root of variance by determining the variation between each data point relative to the mean.
- Standard deviation is the most commonly used measure of dispersion of data around a mean - described more frequently than the variance. Arithmetically, standard deviation is defined as the square root of the variance.
- The Standard Deviation (SD) measures the amount of variability, or dispersion, for a subject set of data from the mean, while the Standard Error of the Mean (SEM) measures how far the sample mean of the data is

likely to be from the true mean. The SEM is always smaller than the SD. Both the 'Standard Deviation' and 'Standard Error (SE)' are often used in experimental studies.

- The meaning of SEM includes statistical inference based on the sampling distribution. SEM is the SD of the theoretical distribution of the sample means (the sampling distribution).
- In probability theory and statistics, the Coefficient of Variation (CV), also known as Relative Standard Deviation (RSD), is a standardized measure of dispersion of a probability distribution or frequency distribution.
- The CV or RSD is widely used in analytical chemistry to express the precision and repeatability of an assay. It is also commonly used in fields, such as engineering or physics when doing quality assurance studies and ANalysis of VAriance (ANOVA) gauge Repeatability and Reproducibility (R&R).

8.5 KEY WORDS

- **Arithmetic mean:** The arithmetic mean or simply mean is the sum of the observations divided by the total number of observations. It is the most common statistic of central tendency which simply gives 'The Mean' or 'The Average' of the dataset.
- **Median:** The median is the middle number obtained by ordering/organizing all data points and picking out the one in the middle or if there are two middle numbers, then taking the mean of those two numbers.
- **Mode:** The mode is referred as the most frequent number occurring in the dataset, i.e., the number that occurs the highest number of times. Therefore the mode is the most commonly occurring data point in a dataset. The mode is useful when there are a lot of repeated values in a dataset.
- **Standard Deviation (SD):** The Standard Deviation (SD) is a statistic that measures the dispersion of a dataset relative to its mean and is calculated as the square root of the variance.
- **Coefficient of Variation (CV):** The Coefficient of Variation (CV) is defined as the ratio of the standard deviation σ to the mean μ, $C_V = \sigma/\mu$.

8.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. What is arithmetic mean?
2. Explain about the calculation of mean for grouped data.
3. Give the merits and demerits of arithmetic mean.

NOTES

4. Elaborate on the median.
5. Explain about the median value for ungrouped data.
6. What do you understand by mode?
7. What is Standard Deviation (SD)?
8. Interpret the Standard Error of the Mean (SEM).
9. Define the term Coefficient of Variation (CV).

Long-Answer Questions

1. Discuss briefly about the arithmetic mean and give the calculation of mean for grouped and ungrouped data.
2. What is median? Explain the median value for grouped and ungrouped data with appropriate examples.
3. Describe the mode with merits, demerits and its uses.
4. Briefly explain about the SD and SEM with the help of examples.
5. Analyse the CV. Give the appropriate examples.

8.7 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 9 PROBABILITY AND HYPOTHESIS TESTING

NOTES

Structure

9.0 INTRODUCTION

Probability is the branch of mathematics which deals with numerical descriptions of how likely an event is to occur, or how likely it is that a proposition is true. Probability and Statistics are the two important concepts in Maths. Probability is all about chance. Whereas statistics is more about how we handle various data using different techniques. It helps to represent complicated data in a very easy and understandable way. A statistical hypothesis is a hypothesis that is testable on the basis of observed data modelled as the realised values taken by a collection of random variables. A statistical hypothesis test is a method of statistical inference.

In probability theory, a normal (or Gaussian or Gauss or Laplace–Gauss) distribution is a type of continuous probability distribution for a real-valued random variable. In statistics, a Confidence Interval (CI) is the means of estimate computed from the statistics of the observed data. This gives a range of values for an unknown parameter (for example, a population mean). The interval has an associated confidence level that gives the probability with which an estimated interval will contain the true value of the parameter. The confidence level is chosen by the investigator. The *p*-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A very small *p*-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.

In this unit, you will study about the normal distribution, confidence interval and *p*-value.

NOTES

9.1 OBJECTIVES

After going through this unit, you will be able to:

- Discuss about the normal distribution, confidence interval and p- value

9.2 NORMAL DISTRIBUTION, CONFIDENCE INTERVAL AND P VALUE

In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events. For example, if the random variable X is used to denote the outcome of a coin toss of the experiment, then the probability distribution of X would take the value 0.5 for X = Heads, and 0.5 for X = Tails assuming that the coin is unbiased. Examples of random phenomena can include the results of an experiment or survey.

Fundamentally, a probability distribution specifies the probability of getting an observation in a particular range of values. Distribution is a significant measure of analysing data sets which indicates all the potential outcomes of the data, and how frequently they occur. The 'Normal Distribution' describes continuous data which have a symmetric distribution, with a characteristic 'Bell-Shaped' curve. The 'Binomial Distribution' describes the distribution of binary data from a finite sample. The 'Poisson Distribution' describes the distribution of binary data from an infinite sample.

Normal Distribution

Normal distribution is often termed as a bell curve and is generally utilized in statistics, business settings, and government entities.

Normal distribution holds the following characteristics:

- It occurs naturally in numerous situations.
- Data points are similar and occur within a small range.
- The mean, mode and median are all equal.
- The curve is symmetric at the centre, i.e., around the mean, μ.
- The curve of the distribution is bell-shaped and symmetrical about the line $x = \mu$.
- The total area under the curve is 1.
- Exactly half of the values are to the left of the centre and the other half to the right.

- Can be utilized to model risks following the distribution of likely outcomes for certain events.
- The formula for calculating the Normal Distribution is,

$$Z = \frac{X - \mu}{\sigma}$$

Where,

X = Value that is being Consistent

μ = Mean of the Distribution

σ = Standard Deviation of the Distribution

A standard normal distribution is defined as the distribution with mean 0 and standard deviation 1 for the PDF (Partial Differential Equation) such that it becomes:

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \qquad for -\infty < x < \infty$$

Figure 9.1 illustrates the curve for standard normal distribution.

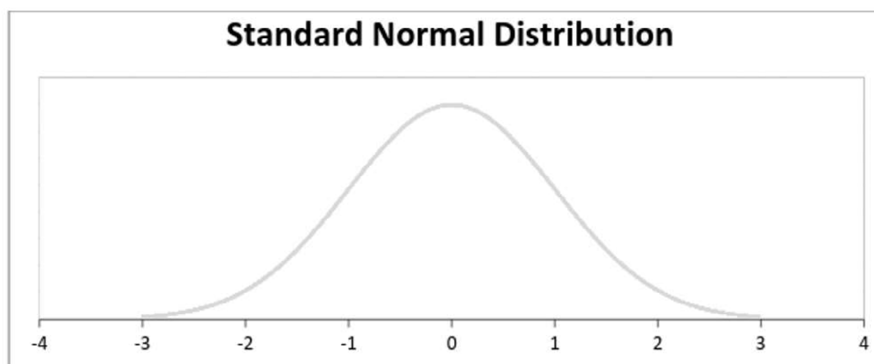


Fig. 9.1 *Curve for Standard Normal Distribution*

The 'Normal Distribution' can be represented by the histogram of a continuous variable obtained from a single measurement on different subjects will have a characteristic 'Bell Shaped' distribution curve termed as the Normal distribution. The normal distribution can be represented as histogram, for example the curve shown in Figure 9.2 represents the birth weight of the 3,226 new born babies (in kilograms).

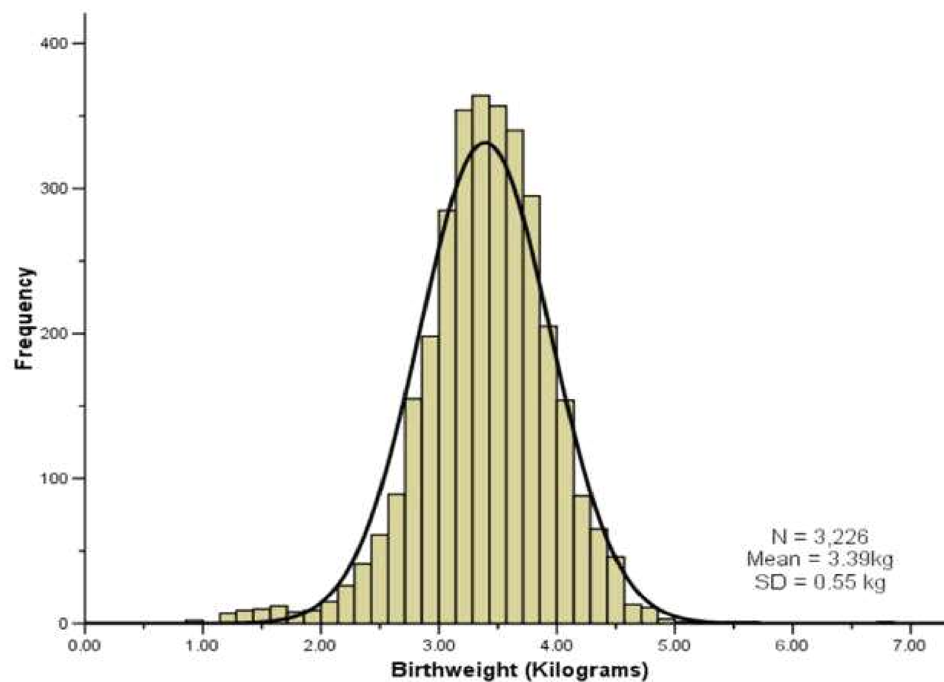


Fig. 9.2 *Histogram showing the Distribution Curve for the Birth Weight of 3,226 New Born Babies (Data from O' Cathain et al. 2002)*

Figure 9.2 shows the histogram of the sample data for an estimate of the population distribution of birth weights in new born babies. This population distribution can be estimated by the superimposed smooth 'Bell Shaped' curve or 'Normal Distribution'. Considering the entire population of new born babies and plotting the histogram of the distribution of birth weight would have exactly the 'Normal Shape'.

The Normal distribution is described by two parameters μ and σ, where μ represents the population mean, or centre of the distribution, and σ the population standard deviation. It is symmetrically distributed around the mean. Populations with small values of the standard deviation σ have a distribution concentrated close to the centre μ, those with large standard deviation have a distribution widely spread along the measurement axis. One mathematical property of the Normal distribution is that exactly 95% of the distribution lies between,

$\mu - (1.96x\sigma)$ and $\mu + (1.96x\sigma)$

Altering the multiplier 1.96 to 2.58, exactly 99% of the Normal distribution lies in the corresponding interval.

In practice the two parameters of the Normal distribution, μ and σ, must be estimated from the sample data. For this a random sample from the population is taken. The sample mean and the sample standard deviation, , are then calculated.

If a sample is taken from such a Normal distribution, and provided the sample is not too small, then approximately 95% of the sample lie within the interval:

$$\bar{x} - \left[1.96 \times SD(\bar{x})\right] \text{ to } \bar{x} + \left[1.96 \times SD(\bar{x})\right]$$

This is calculated by merely replacing the population parameters μ and σ by the sample estimates and S in the previous expression. In the appropriate situations this interval may estimate the reference interval for any required specific laboratory test which can be used for analysis and diagnostic determinations.

To calculate the reference range, consider that the sample birth weight data look normally distributed. As already mentioned that about 95% of the observations from a Normal distribution lie within ±1.96 SDs of the mean. Therefore a reference range for our sample of new born babies, using the values as represented in the histogram of Figure 1.2, is:

= 3.39 – [1.96 × 0.55] to 3.39 + [1.96 × 0.55]

= 2.31 kg to 4.47 kg

Binomial Distribution

Binomial Distribution is considered as the likelihood of a pass or fail outcome in a survey or experiment that is replicated numerous times. There are only two potential outcomes for this type of distribution, such as a True or False, or Heads or Tails. For example, assume that on flipping the coin you won the toss, i.e., Head is appeared, then this indicates a successful event. There are only two possible outcomes. Head denoting success and tail denoting failure. Therefore, probability of getting a Head = 0.5 and the probability of failure, i.e., getting a Tail = 0.5. A distribution where only two outcomes are possible, such as success or failure, gain or loss, win or lose, true or false, and where the probability of success and failure is same for all the trials then it is termed as a Binomial Distribution.

Each trial is independent since the outcome of the previous toss does not determine or affect the outcome of the current toss. An experiment with only two possible outcomes repeated **n** number of times is called binomial. The parameters of a binomial distribution are **n** and **p** where **n** is the total number of trials and **p** is the probability of success in each trial.

The characteristic properties of a Binomial Distribution are:

1. Each trial is independent.
2. There are only two possible outcomes in a trial - either a success or a failure.
3. A total number of **n** identical trials are conducted.
4. The probability of success and failure is same for all trials. Trials are identical.

NOTES

The mathematical representation of binomial distribution is given by:

$$P(x) = \frac{n!}{(n-x)!x!} p^x q^{n-x}$$

A binomial distribution graph where the probability of success does not equal the probability of failure looks as shown in Figure 9.3.

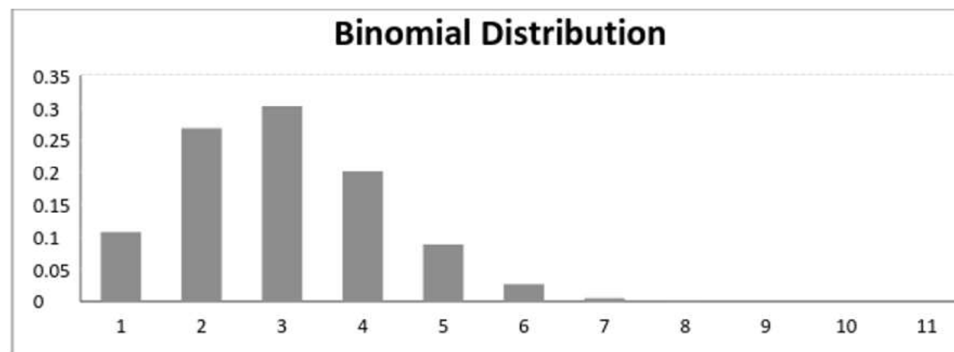


Fig. 9.3 *Binomial Distribution Graph*

Further, when

Probability of Success = Probability of Failure

Then in such a condition the graph of binomial distribution appears as shown in Figure 9.4.

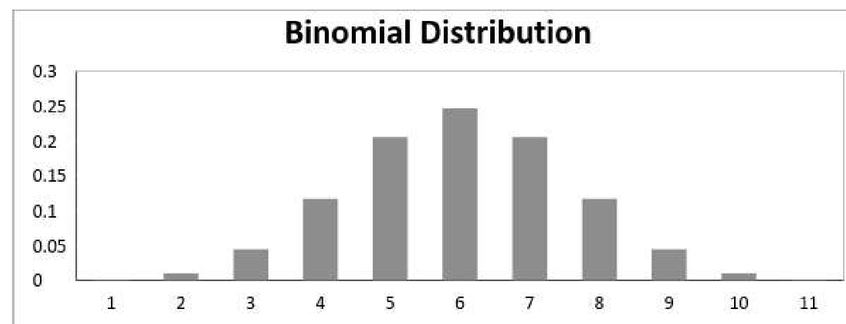


Fig. 9.4 *Binomial Distribution Graph for Probability of Success = Probability of Failure*

The mean and variance of a binomial distribution are given by:

Mean $\rightarrow \mu = n*p$

Variance $\rightarrow$ Var(X) = n*p*q

Poisson Distribution

The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events

occur with a known constant rate and independently of the time since the last event. The Poisson distribution can also be used for the number of events in other specified intervals, such as distance, area or volume. The Poisson distribution is used to describe discrete quantitative data, such as counts in which the population size **n** is large, the probability of an individual event is small, but the expected number of events, **n**, is moderate (say five or more). Typical examples are the number of deaths in a town from a particular disease per day, or the number of admissions to a particular hospital. Poisson distribution is applicable in situations where events occur at random points of time and space, but we will only consider the number of occurrences of the event.

A distribution is called **Poisson distribution** when the following assumptions are valid:

1. Any successful event should not influence the outcome of another successful event.
2. The probability of success over a short interval must equal the probability of success over a longer interval.
3. The probability of success in an interval approaches zero as the interval becomes smaller.

Now, if any distribution validates the above assumptions then it is a Poisson distribution. The notations used in Poisson distribution are:

- λ = Rate at which an event occurs.
- t = Length of a time interval.
- X = Number of events in that time interval.

Here, X is called a 'Poisson Random Variable' and the probability distribution of X is called Poisson distribution. For example, if μ denote the mean number of events in an interval of length t. Then, $\mu = \lambda$*t.

The probability of $X = x$ following a Poisson distribution is given by:

$$P(X = x) = e^{-\mu} \frac{\mu^x}{x!} \qquad \text{for } x = 0, 1, 2, ...,$$

The mean μ is the parameter of this distribution. In addition, the μ is also defined as the λ times length of that interval. The graph of a Poisson distribution is shown in Figure 9.5.

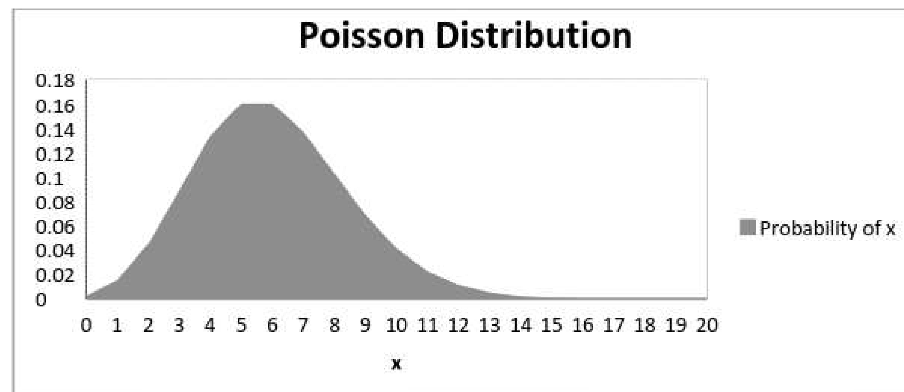


Fig. 9.5 *Poisson Distribution for Probability of* x

The graph shown in Figure 9.6 illustrates the shift in the curve due to increase in the mean.

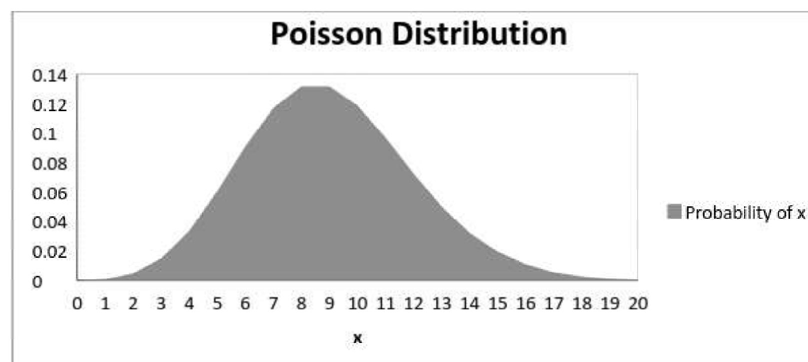


Fig. 9.6 *Graph for the Shift in the Curve due to Increase in Mean*

It is observed that as the mean increases, the curve shifts to the right.

The mean and variance of *X* following a Poisson distribution is given as:

Mean $\rightarrow$ E(X) = μ

Variance $\rightarrow$ Var(X) = μ

9.2.1 Interval Estimation

Point estimator, though simplistic in nature, has some drawbacks. First, a point estimator from the sample may not exactly locate the population parameter resulting in some margin of uncertainty. The average of a sample, for example may or may not be equal or close to the average of the population. If the sample average is different from the population average, the point estimator does not indicate the extent of the possible error, even though this error can be reduced by increasing the sample size. Second, a point estimate does not specify as to how confident we can be that the estimate is close to the parameter it is estimating.

To reasonably overcome these drawbacks, statisticians use another type of estimation known as interval estimation. In this method, we first find a point estimate.

Then we use this estimate to construct an interval on both sides of the point estimate, within which we can be reasonably confident that the true parameter will lie. For example, suppose that we wanted to find out the average salary of full professors at a university who had served at least five years at that rank. Suppose further, that a random sample was taken and the average of the sample was computed to be \$55,000. It is quite possible that the actual average salary of all university professors is \$55,000. However, it is equally possible that the sample was not true representative of the population and the average of the population is quite far off the average of the sample. Accordingly, it is much more likely that the average salary of all the professor's lies somewhere, let us say, between \$50,000 and \$60,000 than exactly at \$55,000. Of course, the greater the range of interval around the sample mean, the more likely it is that the population mean lies in that range. This degree of likelihood is known as the confidence level and the range around the sample mean is known as the confidence interval at a given confidence level. (It is, of course, assumed that the sample is large enough so that the Central Limit Theorem holds.)

Interval Estimate of the Population Mean (Population Variance Known)

Since the sample means are normally distributed, with a mean of μ and a standard deviation of $\sigma_{\bar{X}}$ it follows that sample means follow normal distribution characteristics. Transforming the sampling distribution of sample means into the standard normal distribution, we get:

$$Z = \frac{\bar{X} - \mu}{\sigma_{\bar{X}}}$$

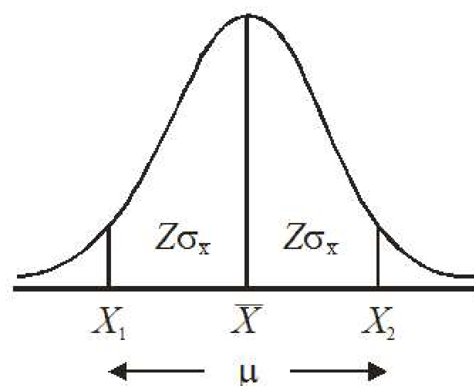
$$or\ \bar{X} - \mu = Z\sigma_{\bar{X}}$$

$$or\ \mu = \bar{X} - Z\sigma_{\bar{X}}$$

Since μ falls within a range of values equidistant from $\bar{X}$,

$$\mu = \bar{X} \pm Z\sigma_{\bar{X}}$$

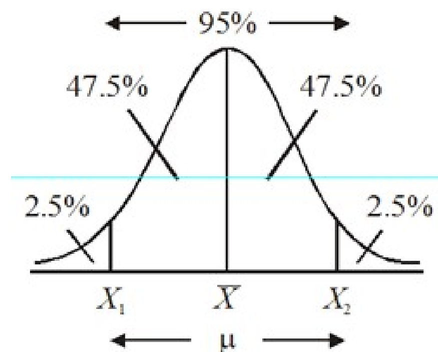
This relationship is shown in the following illustration.



NOTES

This means that the population mean is expected to lie between the values of X_1 and X_2 which are both equidistant from and this distance depends upon the value of Z which is a function of confidence level.

Suppose that we wanted to find out a confidence interval around the sample mean within which the population mean is expected to lie 95 percent of the time. (We can never be sure that the population mean will lie in any given interval 100 percent of the time). This confidence interval is shown as follows:



The points X_1 and X_2 above define the range of the confidence interval as follows:

$$X_1 = \overline{X} - Z\sigma_{\overline{X}}$$
$$\text{and } X_2 = \overline{X} + Z\sigma_{\overline{X}}$$

Looking at the table of Z scores, (given in the Appendix) we find that the value of Z score for area 10.4750 (half of 95 per cent) is 1.96. This illustration can be interpreted as follows:

(a) If all possible samples of size n were taken, then on the average 95 per cent of these samples would include the population mean within the interval around their sample means bounded by X_1 and X_2.

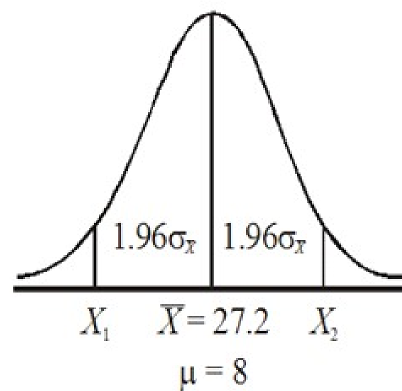
(b) If we took a random sample of size n from a given population, the probability is 0.95 that the population mean would lie between the interval X_1 and X_2 around the sample mean, as shown.

(c) If a random sample of size n was taken from a given population, we can be 95 percent confident in our assertion that the population mean will lie around the sample mean in the interval bounded by values of X_1 and X_2 as shown. (It is also known as 95 per cent confidence interval.) At 95 per cent confidence interval, the value of Z score as taken from the Z score table is 1.96. The value of Z score can be found for any given level of confidence, but generally speaking, a confidence level of 90 per cent, 95 per cent or 99 per cent is taken into consideration for which the Z score values are 1.68, 1.96 and 2.58, respectively.

Example 9.1: The sponsor of a television programme targeted at the children's market (age 4-10 years) wants to find out the average amount of time children

spend watching television. A random sample of 100 children indicated the average time spent by these children watching television per week to be 27.2 hours. From previous experience, the population standard deviation of the weekly extent of television watched (σ) is known to be 8 hours. A confidence level of 95 percent is considered to be adequate.

Solution:



The confidence interval is given by:

$$\overline{X} \pm Z\sigma_{\overline{X}} \qquad \text{or } \overline{X} - Z\sigma_{\overline{X}} < \mu < \overline{X} + Z\sigma_{\overline{X}}$$

Where : $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}}$

Accordingly, we need only four values, namely, *Z, s* and *n.* In our case:

$$\overline{X} = 27.2$$
$$Z = 1.96$$
$$\sigma = 8$$
$$n = 100$$

Hence $\sigma_{\overline{X}} = \dfrac{\sigma}{\sqrt{n}} = \dfrac{8}{\sqrt{100}} = \dfrac{8}{10} = 0.8$

Then,

$$X_1 = \overline{X} - Z\sigma_{\overline{X}}$$
$$= 27.2 - (1.96 \times 0.8) = 27.2 - 1.568$$
$$= 25.632$$

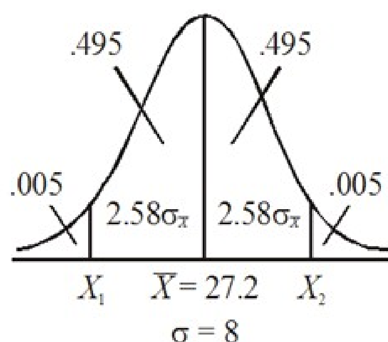
and

$$X_2 = \overline{X} + Z\sigma_{\overline{X}}$$
$$= 27.2 + (1.96 \times 0.8) = 27.2 + 1.568$$
$$= 28.768$$

This means that we can conclude with 95 percent confidence that a child on an average spends between 25.632 and 28.768 hours per week watching television. (It should be understood that 5 percent of the time our conclusion would still be wrong. This means that because of the symmetry of distribution, we will be wrong 2.5 percent of the times because the children on an average would be watching television more than 28.768 hours and another 2.5 percent of the time we will be wrong in our conclusion, because on an average, the children will be watching television less than 25.632 hours per week.)

Example 9.2: Calculate the confidence interval in the previous problem, if we want to increase our confidence level from 95 per cent to 99 per cent. Other values remain the same.

Solution:



If we increase our confidence level to 99 percent, then it would be natural to assume that the range of the confidence interval would be wider, because we would want to include more values which may be greater than 28.768 or smaller than 25.632 within the confidence interval range. Accordingly, in this new situation,

$$Z = 2.58$$

$$\sigma_{\bar{X}} = 0.8$$

Then

$$X_1 = \bar{X} - Z\sigma_{\bar{X}}$$
$$= 27.2 - (2.58\ X\ 0.8) = 27.2 - 2.064$$
$$= 25.136$$

And

$$X_2 = \bar{X} + Z\sigma_{\bar{X}}$$
$$= 27.2 + 2.064$$
$$= 29.264$$

NOTES

(The value of Z is established from the table of Z scores against the area of 0.495 or a figure closest to it. The table shows that the area close to 0.495 is 0.4949 for which the Z score is 2.57 or 0.4951 for which the Z score is 2.58. In practice, the Z score of 2.58 is taken into consideration when calculating 99 percent confidence interval.)

Interval Estimate of the Population Mean (Population Variance Unknown)

As the previous example shows, in order to determine the interval estimate of μ, the variance (and hence, the standard deviation) must be known, since it figures in the formula. However, the standard deviation of the population is generally not known. In such situations and when sample size is reasonably large (30 or more), we can approximate the population standard deviation (σ) by the sample standard deviation (σ), so that the confidence interval,

$$\overline{X} \pm Z\sigma_{\overline{X}} \quad \text{is approximated by the interval}$$

$$X \pm Zs_{\overline{X}}, \text{ when } n \geq 30.$$

$$\text{where } \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \text{ and } s_{\bar{X}} = \frac{s}{\sqrt{n}}$$

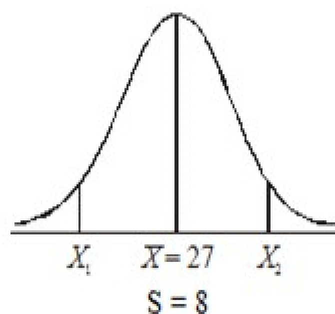
Example 9.3: It is desired to estimate the average age of students who graduate with an MBA degree in the university system. A random sample of 64 graduating students showed that the average age was 27 years with a standard deviation of 4 years.

(a) Estimate a 95 per cent confidence interval estimate of the true average (population mean) age of all such graduating students at the university.

(b) The confidence interval limits change if the confidence level was increased from 95 percent to 99 percent.

Solution: Since the sample size n is sufficiently large, we can approximate the population standard deviation by the sample standard deviation.

(a)



Now,

$$Z = 1.96$$
$$\overline{X} = 27$$
$$s_{\overline{x}} = \frac{s}{\sqrt{n}} = \frac{4}{\sqrt{64}} = \frac{4}{8} = 0.5$$

95 percent confidence interval of population mean μ, is given by:

$$\overline{X} \pm Zs_{\overline{X}}$$

So that,

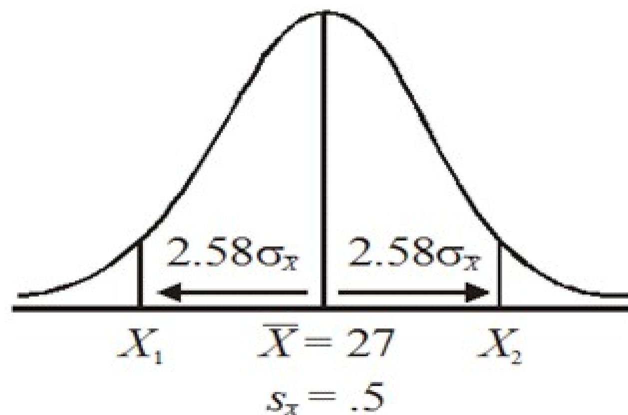
$$X_1 = \overline{X} - Zs_{\overline{X}}$$
$$= 27 - (1.96 \times 0.5) = 27 - 0.98$$
$$= 26.02$$

And,

$$X_2 = \overline{X} + Zs_{\overline{X}}$$
$$= 27 + 0.98$$
$$= 27.98$$

Hence, $26.02 \geq \mu \geq 27.98$.

(b) $s_{\overline{X}} = 0.5$



NOTES

Now, Z becomes 2.58 and the other values remain the same. Hence,

$$X_1 = \overline{X} - Zs_{\overline{X}}$$
$$= 27 - (2.58 \times 0.5) = 27 - 1.29$$
$$= 25.71$$

And,

$$X_2 = \overline{X} + Zs_{\overline{X}}$$
$$= 27 + 1.29$$
$$= 28.29$$

Hence, $25.71 \leq \mu \leq 28.29$.

Sample Size Determination for Estimating the Population Mean

It is understood that the larger the sample size, the closer the sample statistic will be to the population parameter. Hence, the degree of accuracy we require in our estimate would be one factor influencing our choice of sample size. The second element that influences the choice of the sample size is the degree of confidence in ourselves that the error in the estimate remains within the degree of accuracy that is desired. Hence, the degree of accuracy has two aspects.

1. The maximum allowable error in our estimate
2. The degree of confidence that the error in our estimate will not exceed the maximum allowable error

The ideal situation would be that the sample mean $\overline{X}$ equals the population mean μ. That would be the best estimate of μ based on $\overline{X}$. If the entire population was taken as a sample then $\overline{X}$ will be equal to μ and there will be no error in our estimate. Hence, ($\overline{X} - \mu$) can be considered as *error* or *deviation* of the estimator from the population mean μ. This maximum allowable *error* must be pre-established. Let this error be denoted by E, so that:

$$E = (\overline{X} - \mu)$$

Now, we know that,

$$Z = \frac{\overline{X} - \mu}{\sigma_{\overline{X}}}$$
$$= \frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$
$$= \frac{E}{\sigma / \sqrt{n}}$$

NOTES

Or

$$Z = \frac{E\sqrt{n}}{\sigma}$$

$$Z\sigma = E\sqrt{n}$$

$$\sqrt{n} = \frac{Z\sigma}{E}$$

$$n = \left(\frac{Z\sigma}{E}\right)^2 \text{ or } \frac{Z^2\sigma^2}{E^2}$$

Based upon this formula, it can be seen that the size of the sample depends upon:

(a) Confidence interval desired. This will determine the value of Z. For example, 95 per cent confidence level yields the value of Z to be = 1.96.

(b) Maximum error allowed (*E*).

(c) The standard deviation of the population (σ).

It can further be seen from this formula that the sample size will increase if:

(a) The allowable error becomes smaller.

(b) The degree of confidence increases.

(c) The value of the variance within the population is larger.

Example 9.4: We would like to know the average time that a child spends watching television over the weekend. We want our estimate to be within 1 hour of the true population average. (This means that the maximum allowable error is 1 hour.) Previous studies have shown the population standard deviation to be 3 hours. What sample size should be taken for this purpose, if we want to be 95 percent confident that the error in our estimate will not exceed the maximum allowable error?

Solution: For 95 per cent confidence level, the values of

Z = 1.96

E = 1 hour (given)

σ = 3 hours (given)

Then,

$$n = \frac{Z^2\sigma^2}{E^2}$$

$$= \frac{(1.96)^2(3)^2}{(1)^2}$$

$$= 34.57$$

To be more accurate in our estimate, we always round off the answer to the next higher figure from the decimal. Hence, $n = 35$.

Confidence Interval Estimation of Population Proportion

So far we have discussed the estimation of population mean, which is quantitative in nature. This concept of estimation can be extended to qualitative data where the data is available in proportion or percentage form. In this situation the parameter of interest is π, which is the proportion of times a certain desirable outcome occurs. This concept lends itself to binomial distribution where we label the outcome of interest to us as success with the probability of success being π and the probability of failure being $(1-\pi)$.

When large samples of size n are selected from a population having a proportion of desirable outcomes π, then the sampling distribution of proportions is normally distributed. For large samples, when (np) as well as (nq) are both at least equal to 5, where n is the sample size, p is the probability of a desired outcome (or success) and q is the probability of failure $(1–p)$, then the binomial distribution can also be approximated to normal distribution, with a mean of π and a standard deviation of σ_p, where σ_p is given by:

$$\sigma_p = \sqrt{\frac{\pi(1-\pi)}{n}}$$

In such cases, we expect 95 per cent of all sample proportions to fall within the following range:

$$\pi \pm 1.96\sigma_p$$

Or,

$$\pi \pm 1.96\sqrt{\frac{\pi(1-\pi)}{n}}$$

Similarly, 99 per cent of all such sample proportions will fall within

$$\pi \pm 2.58\sqrt{\frac{\pi(1-\pi)}{n}}$$

If all possible samples of size n are selected and the interval is established for each sample, where p is the sample proportion, then 95 percent of all such intervals are expected to contain π, the population proportion. Then this range of

$$p \pm 1.96\sigma_p$$

$$p \pm 1.96\sqrt{\frac{\pi(1-\pi)}{n}}$$

Now 99 percent confidence interval estimate of π becomes:

$$p \pm 2.58\ \sigma_p$$

Example 9.5: A survey of 500 persons shopping at a mall, selected at random, showed that 350 of them used credit cards for their purchases and 150 of them used cash.

(a) Construct a 95 percent confidence interval estimate of the proportion of all persons at the mall, who use credit card for shopping.

(b) What would our confidence level be, if we make the assertion that the proportion of shoppers at the mall who shop with a credit card is between 67 percent and 73 per cent.

Solution:

(a) There are 350 people out of a total sample of 500 who pay by credit card. Hence, the sample proportion of credit card shoppers is:

p = 350/500 = 0.7

The 95 percent confidence interval estimate of population proportion π is given as follows:

$$p \pm 1.96\ \sigma_p$$

Where,

$$\sigma_p = \sqrt{\frac{p(1-p)}{n}}$$

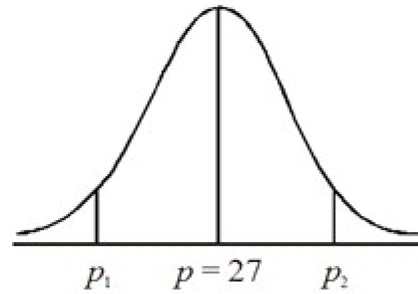
(Since π is not known, we approximate sample proportion p for population proportion π).

Then, $$\sigma_p = \sqrt{\frac{0.7(0.3)}{500}} = \sqrt{0.00042} = 0.02$$

Then the confidence limits are:

$$p_1 = p - 1.96\ \sigma_p$$
$$= 0.7 - 1.96(0.02)$$
$$= 0.7 - 0.0392 = 0.6608 \text{ or } 66.08\%, \text{ and}$$
$$p_2 = p + 1.96\ \sigma_p$$
$$= 0.7 + 0.0392 = 0.7392 \text{ or } 73.92\%$$

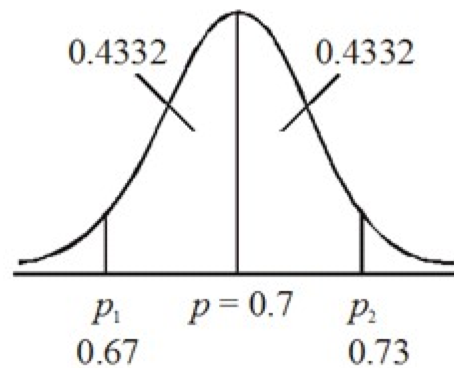
NOTES



0.6608 $\quad \sigma_p = 0.02 \quad$ 0.7392

This means that the population of people who pay by credit card at the mall is between 66.8 percent and 73.92 percent.

(b) If the population proportion of credit card shoppers is given to be between 0.67 and 0.73, when such sample proportion p is 0.70, then



$$p_1 = p - Z\sigma_p$$
$$0.67 = 0.7 - Z\ (0.2)$$
$$0.02 = 0.7 - 0.67$$
$$Z = \frac{0.7 - 0.67}{0.02} = \frac{0.03}{0.02} = 1.5$$

Similarly,

$$p_2 = p + Z\sigma_p$$
$$0.73 = 0.7 + Z\ (0.02)$$
$$Z = \frac{0.73 - 0.7}{0.02} = \frac{0.03}{0.02} = 1.5$$

NOTES

Using the Z score table, we see that the area under the curve for Z=1.5 is 0.4332. This area is on each side of the mean so that the total area is 0.8664. In other words, our confidence level is 86.64 percent that the proportion of shoppers using credit card is between 67 per cent and 73 percent.

Sample Size Determination for Estimating the Population Proportion

We follow the same procedure as we did in determining the sample size for estimating the population mean. As before, there are three factors that are taken into consideration. These are:

(a) The level of confidence desired.

(b) The maximum allowable error permitted in the estimate.

(c) The estimated population proportion of success p.

As established previously,

$$Z = \frac{p-\pi}{\sigma_p}$$

Where, $\sigma_p = \sqrt{\frac{\pi(1-\pi)}{\sigma_p}}$

Now, $(p-\pi)$ can be considered as error (E), so that:

$$Z = \frac{E}{\sqrt{\frac{\pi(1-\pi)}{n}}}$$

By cross-multiplication we get,

$$E = Z\sqrt{\frac{\pi(1-\pi)}{n}}$$

Squaring both sides we get,

$$E^2 = \frac{Z^2\pi(1-\pi)}{n}$$

$$\text{or } nE^2 = Z^2\pi(1-\pi)$$

$$\text{or } n = \frac{Z^2\pi(1-\pi)}{E^2}$$

This formula assumes that we know π, the population proportion, which we are trying to estimate in the first place. Accordingly, π is unknown. However, if any previous studies have estimated this value or a sample proportion p has

been calculated in previous studies, then we can approximate this p for π and hence,

$$n = \frac{Z^2 p(1-p)}{E^2}$$

However, if no previous surveys have been taken so that we do not know the value of π or p, then we assume π to be equal to 0.5, simply because, other things being given, the value of π being 0.5 will result in a larger sample size than any other value assumed by π. Hence, the sample size would be at least as large as or larger than required for the given conditions. This can be established by the fact that when $\pi = 0.5$, then $\pi(1-\pi)$ is $0.5 \times 0.5 = 0.25$. This value is larger than any other value of $\pi(1-\pi)$. This means that when p = 0.5 then for a given value of Z, n would be larger than any other value of π. This results in a more conservative estimate which is desirable.

Example 9.6: It is desired to estimate the proportion of children watching television on Saturday mornings, in order to develop a promotional strategy for electronic games. We want to be 95 percent confident that our estimate will be within ± 2 percent of the true population proportion.

(a) What sample size should we take if a previous survey showed that 40 per cent of children watched television on Saturday mornings?

(b) What would be the sample size, for the same degree of confidence and the same maximum allowable error, if no such previous survey had been taken?

Solution: (a) In this case, the following values are given:

$Z = 1.96$ (95% Confidence Interval)

$p = 0.4$

$E = 0.02$

Substituting these values in the following formula, we get:

$$n = \frac{Z^2 p(1-p)}{E^2}$$

$$= \frac{(1.96)^2(0.4)(0.6)}{(0.02)^2}$$

$$= \frac{0.922}{0.0004}$$

$$= 2304.96$$

$$= 2305 \text{ (approx.)}$$

For the sake of accuracy, we always round off to the next higher figure, in case of answer being a fraction.

NOTES

(b) In this case, since no previous surveys have been taken, we assume p = 0.5 and follow the earlier procedure.

$$n = \frac{Z^2 p(1-p)}{E^2}$$

$$= \frac{(1.96)^2 (0.5)(0.5)}{(0.02)^2}$$

$$= \frac{0.9604}{0.0004}$$

$$= 2401$$

Confidence Interval of Variance

Confidence Interval (CI) tells us about an interval in which estimates of a population parameter can be reliably predicted. Confidence intervals and estimates have applicability for a whole range of quantitative studies. A single value is not used for estimating the parameter; rather an interval is chosen that is likely to include the parameter. Hence, confidence interval indicates the reliability of an estimate. **Confidence level or confidence coefficient** is the likelihood of an interval to contain the parameter. The desired confidence level can be increased by widening the confidence interval.

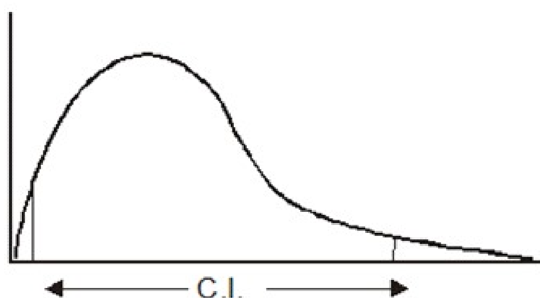
We can select a confidence interval to indicate the reliability of a survey result. In an opinion poll, results might indicate that 40 per cent of the people who participated may vote for a particular party. To get a 95 percent confidence interval, we say that 95 percent of those who participated had the same opinion on the date of the survey. Keeping other things equal, the result with a small CI is found more reliable than that with a large CI. One factor which controls this width, as in this case, is the sample size.

The test leading to a statistically significant claim with a confidence interval is performed at a significance level of $(1-\alpha).100$ percent. If p be the confidence level for a given proportion, a confidence interval for this parameter is an interval calculated from a random sample of an underlying population. This sample is such that in the event of repetition of the sampling with the confidence interval recalculated with the same method, for each sample a proportion of the confidence intervals would contain the same population parameter. Sometimes a confidence set may consist of several separate intervals and may include intervals which are semi-infinite. It may be possible that the outcome of such a calculation is the set of values lying in the set of real numbers which are from minus infinity to plus infinity, i.e., $(-\infty, \infty)$.

Interval estimates may show a contrast with point estimates and indicate the precision of the parameter. In practice, 95 percent confidence level is assumed for finding a confidence interval. In a graphical presentation, confidence intervals can be based on several confidence levels, e.g., 50 per cent, 95 percent or 99 percent.

Chi-Squared Distribution for Confidence Interval of Variance

Chi-squared distribution interval of variance can also be evaluated for a Chi-squared distribution. The following example will make the concept clear. The area under the curve, between the critical values and excluding the tail areas is known as α risk. It gives the confidence interval. The most likely distribution of population variances is indicated by the whole region of the curve for a given sample size and variation. Here, the population variation, σ is approximated by sample variance, S for simplification.



$$\text{C.I} = \frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}$$

Where,

n = The sample size

S_2 = The sample variance

$\frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}}$ & $\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,\,n-1}}$ = the χ^2 distribution values for the desired confidence level α and for n–1

Now compute a 95 percent C.I. on variance for a sample (n = 35) with a sample variance, S of 2.3.

$$\text{C.I.} = \frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{\alpha/2,\,n-1}}$$

$$\text{C.I.} = \frac{(34)2.3^2}{51.966} \le \sigma^2 \le \frac{(34)2.3^2}{19.806}$$

$$= 3.5 \text{ to } 9.1$$

This indicates the maximum likelihood of distribution of population variances for a given sample size and variance. This shows that 95 percent of the time, the population's variance is likely to fall in this interval.

NOTES

9.2.2 *p*-Value

In null hypothesis significance testing, the *p*-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A very small *p*-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. Reporting *p*-values of statistical tests is common practice in academic publications of many quantitative fields. Since the precise meaning of *p*-value is hard to grasp, misuse is widespread and has been a major topic in metascience.

Basic Concepts of *p*-Value

In statistics, every conjecture concerning the unknown probability distribution of a collection of random variables representing the observed data ***X*** in some study is called a **statistical hypothesis**. If we state one hypothesis only and the aim of the statistical test is to see whether this hypothesis is tenable, but not, at the same time, to investigate other hypotheses, then such a test is called a **significance test**. Note that the hypothesis might specify the probability distribution of ***X*** precisely, or it might only specify that it belongs to some class of distributions. Often, we reduce the data to a single numerical statistic ***T*** whose marginal probability distribution is closely connected to a main question of interest in the study.

The *p*-value is used in the context of null hypothesis testing in order to quantify the idea of statistical significance of evidence, the evidence being the observed value of the chosen statistic ***T***.

Thus, the only hypothesis that needs to be specified in this test and which embodies the counterclaim is referred to as the null hypothesis; that is, the hypothesis to be nullified. A result is said to be statistically significant if it allows us to reject the null hypothesis. The result, being statistically significant, was highly improbable if the null hypothesis is assumed to be true. A rejection of the null hypothesis implies that the correct hypothesis lies in the logical complement of the null hypothesis. But no specific alternatives need to have been specified. The rejection of the null hypothesis does not tell us which of any possible alternatives might be better supported. However, the user of the test chose the test statistic ***T*** in the first place probably with particular alternatives in mind; such a test is often used precisely in order to convince people that those alternatives are viable because what was actually observed was extremely unlikely under the null hypothesis.

As a particular example, if a null hypothesis states that a certain summary statistic ***T*** follows the standard normal distribution N (0,1), then the rejection of this null hypothesis could mean that

(i) The mean is not 0

(ii) The variance is not 1

(iii) The distribution is not normal.

Different tests of the same null hypothesis would be more or less sensitive to different alternatives. Anyhow, if we do manage to reject the null hypothesis,

NOTES

even if we know the distribution is normal and variance is 1, the null hypothesis test does not tell us which non-zero values of the mean are now most plausible. If one has a huge amount of independent observations from the same probability distribution, one will eventually be able to show that their mean value is not precisely equal to zero; but the deviation from zero could be so small as to have no practical or scientific interest. All other things being equal, smaller *p*-values are taken as stronger evidence against the null hypothesis.

Definition and Interpretation

Consider an observed test-statistic *t* from unknown distribution ***T***. Then the *p*-value *p* is what the prior probability would be of observing a test-statistic value at least as 'Extreme' as *t* if null hypothesis H_0 were true.

If the *p*-value is very small, then either the null hypothesis is false or something unlikely has occurred. In a formal significance test, the null hypothesis H_0 is rejected if the *p*-value is less than a pre-defined threshold value α, which is referred to as the alpha level or significance level. The value of α is instead set by the researcher before examining the data. By convention α is commonly set to 0.05, though lower alpha levels are sometimes used.

The *p*-value is a function of the chosen test statistic *T* and is therefore a random variable. If the null hypothesis fixes the probability distribution of ***T*** precisely, and if that distribution is continuous, then when the null-hypothesis is true, the *p*-value is uniformly distributed between 0 and 1. Thus, the *p*-value is not fixed. If the same test is repeated independently with fresh data (always with the same probability distribution), one will obtain a different *p*-value in each iteration. If the null-hypothesis is composite, or the distribution of the statistic is discrete, the probability of obtaining a *p*-value less than or equal to any number between 0 and 1 is less than or equal to that number, if the null-hypothesis is true. It remains the case that very small values are relatively unlikely if the null-hypothesis is true, and that a significance test at level α is obtained by rejecting the null-hypothesis if the significance level is less than or equal to α.

Different *p*-values based on independent sets of data can be combined, for instance using Fisher's combined probability test.

Usage of *p*-Values

The *p*-value is widely used in statistical hypothesis testing, specifically in null hypothesis significance testing. In this method, as part of experimental design, before performing the experiment, one first chooses a model (the null hypothesis) and a threshold value for *p*, called the significance level of the test, traditionally 5% or 1% and denoted as α. If the *p*-value is less than the chosen significance level (α), that suggests that the observed data is sufficiently inconsistent with the null hypothesis and that the null hypothesis may be rejected. However, that does not prove that the tested hypothesis is false. When the *p*-value is calculated correctly, this test guarantees that the type I error rate is at most α. For typical analysis, using the standard $\alpha = 0.05$ cut-off, the null hypothesis is rejected when *p* d" 0.05 and

NOTES

not rejected when p >0 .05. The p-value does not, in itself, support reasoning about the probabilities of hypotheses but is only a tool for deciding whether to reject the null hypothesis.

Misuse of *p*-Values

According to the American Statistical Association (ASA), there is widespread agreement that p-values are often misused and misinterpreted. One practice that has been particularly criticized is accepting the alternative hypothesis for any p-value nominally less than 0.05 without other supporting evidence. Although p-values are helpful in assessing how incompatible the data are with a specified statistical model, contextual factors must also be considered, such as 'The design of a study, the quality of the measurements, the external evidence for the phenomenon under study, and the validity of assumptions that underlie the data analysis'. Another concern is that the p-value is often misunderstood as being the probability that the null hypothesis is true.

Some statisticians have proposed replacing p-values with alternative measures of evidence, such as confidence intervals, likelihood ratios, or Bayes factors, but there is heated debate on the feasibility of these alternatives. Others have suggested to remove fixed significance thresholds and to interpret p-values as continuous indices of the strength of evidence against the null hypothesis. Yet others suggested to report alongside p-values the prior probability of a real effect that would be required to obtain a false positive risk (i.e., the probability that there is no real effect) below a pre-specified threshold (e.g. 5%).

Check Your Progress

1. Explain about the probability distribution.
2. Elaborate on the normal distribution.
3. When a distribution is called Poisson distribution?
4. How is interval estimation done?
5. What are the aspects of degree of accuracy?
6. Which factors are taken into consideration while determining the sample size for population proportion estimation?
7. What is p-value?

9.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

2. The Normal distribution is described by two parameters μ and σ, where μ represents the population mean, or centre of the distribution, and σ the population standard deviation. It is symmetrically distributed around the mean. Populations with small values of the standard deviation σ have a distribution concentrated close to the centre μ, those with large standard deviation have a distribution widely spread along the measurement axis.
3. A distribution is called Poisson distribution when the following assumptions are valid:
   - Any successful event should not influence the outcome of another successful event.
   - The probability of success over a short interval must equal the probability of success over a longer interval.
   - The probability of success in an interval approaches zero as the interval becomes smaller.
4. An estimator is considered to be efficient if its value remains stable from sample to sample. The best estimator would be the one which would have the least variance from sample to sample taken randomly from the same population. An estimator is said to be sufficient if it uses all the information about the population parameter contained in the sample.
5. The degree of accuracy has two aspects:
   a) The maximum allowable error in our estimate
   b) The degree of confidence that the error in our estimate will not exceed the maximum allowable error.
6. Three factors that are taken into consideration. These are:
   a) The level of confidence desired
   b) The maximum allowable error permitted in the estimate
   c) The estimated population proportion of success p
7. The p-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A very small p-value means that such an extreme observed outcome would be very unlikely under the null hypothesis. Reporting p-values of statistical tests is common practice in academic publications of many quantitative fields.

9.4 SUMMARY

- In probability theory and statistics, a probability distribution is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment.

- A probability distribution specifies the probability of getting an observation in a particular range of values. Distribution is a significant measure of analysing data sets which indicates all the potential outcomes of the data, and how frequently they occur.
- The 'Normal Distribution' describes continuous data which have a symmetric distribution, with a characteristic 'Bell-Shaped' curve. The 'Binomial Distribution' describes the distribution of binary data from a finite sample. The 'Poisson Distribution' describes the distribution of binary data from an infinite sample.
- The 'Normal Distribution' can be represented by the histogram of a continuous variable obtained from a single measurement on different subjects will have a characteristic 'Bell Shaped' distribution curve termed as the Normal distribution.
- In practice the two parameters of the Normal distribution σ and μ, must be estimated from the sample data. For this a random sample from the population is taken. The sample mean and the sample standard deviation, are then calculated.
- Binomial Distribution is considered as the likelihood of a pass or fail outcome in a survey or experiment that is replicated numerous times. There are only two potential outcomes for this type of distribution, such as a True or False, or Heads or Tails.
- The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.
- The Poisson distribution can also be used for the number of events in other specified intervals, such as distance, area or volume.
- Point estimator, though simplistic in nature, has some drawbacks. First, a point estimator from the sample may not exactly locate the population parameter resulting in some margin of uncertainty.
- To reasonably overcome these drawbacks, statisticians use another type of estimation known as interval estimation. In this method, we first find a point estimate. Then we use this estimate to construct an interval on both sides of the point estimate, within which we can be reasonably confident that the true parameter will lie.
- Since the sample means are normally distributed, with a mean of μ and a standard deviation of it follows that sample means follow normal distribution characteristics.

- It is understood that the larger the sample size, the closer the sample statistic will be to the population parameter. Hence, the degree of accuracy we require in our estimate would be one factor influencing our choice of sample size.
- The second element that influences the choice of the sample size is the degree of confidence in ourselves that the error in the estimate remains within the degree of accuracy that is desired.
- Confidence Interval (CI) tells us about an interval in which estimates of a population parameter can be reliably predicted. Confidence intervals and estimates have applicability for a whole range of quantitative studies.
- A single value is not used for estimating the parameter; rather an interval is chosen that is likely to include the parameter. Hence, confidence interval indicates the reliability of an estimate. Confidence level or confidence coefficient is the likelihood of an interval to contain the parameter. The desired confidence level can be increased by widening the confidence interval.
- Chi-squared distribution interval of variance can also be evaluated for a Chi-squared distribution. The following example will make the concept clear. The area under the curve, between the critical values and excluding the tail areas is known as α risk. It gives the confidence interval.
- In null hypothesis significance testing, the *p*-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct. A very small *p*-value means that such an extreme observed outcome would be very unlikely under the null hypothesis.
- Reporting *p*-values of statistical tests is common practice in academic publications of many quantitative fields. Since the precise meaning of *p*-value is hard to grasp, misuse is widespread and has been a major topic in metascience.

9.5 KEY WORDS

- **Probability distribution:** It is a mathematical function that provides the probabilities of occurrence of different possible outcomes in an experiment. In more technical terms, the probability distribution is a description of a random phenomenon in terms of the probabilities of events.
- **Binomial distribution:** Binomial Distribution is considered as the likelihood of a pass or fail outcome in a survey or experiment that is replicated numerous times. There are only two potential outcomes for this type of distribution, such as a True or False, or Heads or Tails.
- **Poisson distribution:** The Poisson distribution is named after the French mathematician Siméon Denis Poisson. It is a discrete probability distribution

NOTES

that expresses the probability of a given number of events occurring in a fixed interval of time or space if these events occur with a known constant rate and independently of the time since the last event.

- **Confidence Interval (CI):** Confidence Interval (CI) tells us about an interval in which estimates of a population parameter can be reliably predicted. Confidence intervals and estimates have applicability for a whole range of quantitative studies.
- ***p*-value:** In null hypothesis significance testing, the *p*-value is the probability of obtaining test results at least as extreme as the results actually observed, under the assumption that the null hypothesis is correct.

9.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the probability distribution.
2. Give the definitions of normal distribution.
3. Elaborate on the binomial distribution.
4. Interpret the poisson distribution.
5. What do you mean by interval estimation?
6. Define the confidence interval.
7. Explain the confidence interval estimation of population proportion.

Long-Answer Questions

1. Discuss briefly about the normal distribution with the help of examples.
2. Analyse the binomial and poission distribution.
3. What is interval estimation? Describe the interval estimation of the population mean for known and unknown population variance.
4. Explain in detail about the confidence interval with appropriate examples.
5. Briefly discuss about the *p*-value giving uses and misuses.

9.7 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 10 COMMON STATISTICAL TOOLS

NOTES

Structure

10.0 INTRODUCTION

A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories. A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t*- distribution.

ANalysis Of VAriance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'Variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalises the *t*-test to more than two groups. ANOVA is useful for comparing/testing three or more group means for statistical significance.

Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. Fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Statistical Package for the Social Sciences (SPSS) is a computer program used for statistical analysis. The features of SPSS incorporate modules for statistical data analysis, including descriptive statistics, such as plots, frequencies, charts and lists, as well as sophisticated inferential and multivariate statistical procedures, such as ANalysis of VAriance (ANOVA), factor analysis, cluster analysis and categorical data analysis.

In this unit, you will study about the chi-square, *t*-test, ANOVA, correlation and regression analysis, statistical packages.

10.1 OBJECTIVES

After going through this unit, you will be able to:

- Define the chi-square
- Understand the *t*-test
- Explain about the ANOVA
- Elaborate on the correlation and regression analysis
- Interpret the statistical packages

10.2 CHI-SQUARE

A **Chi-Squared Test**, also written as χ^2 **test**, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/ categories.

In the standard applications of this test, the observations are classified into mutually exclusive classes, and there is some theory, or say null hypothesis, which gives the probability that any observation falls into the corresponding class. The

NOTES

purpose of the test is to evaluate how likely the observations that are made would be, assuming the null hypothesis is true.

Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.

Also considered a chi-squared test is a test in which this is asymptotically true, meaning that the sampling distribution (if the null hypothesis is true) can be made to approximate a chi-squared distribution as closely as desired by making the sample size large enough. Figure 10.1 illustrates the Chi-squared distribution, showing χ^2 on the x-axis and p-value on the y-axis.

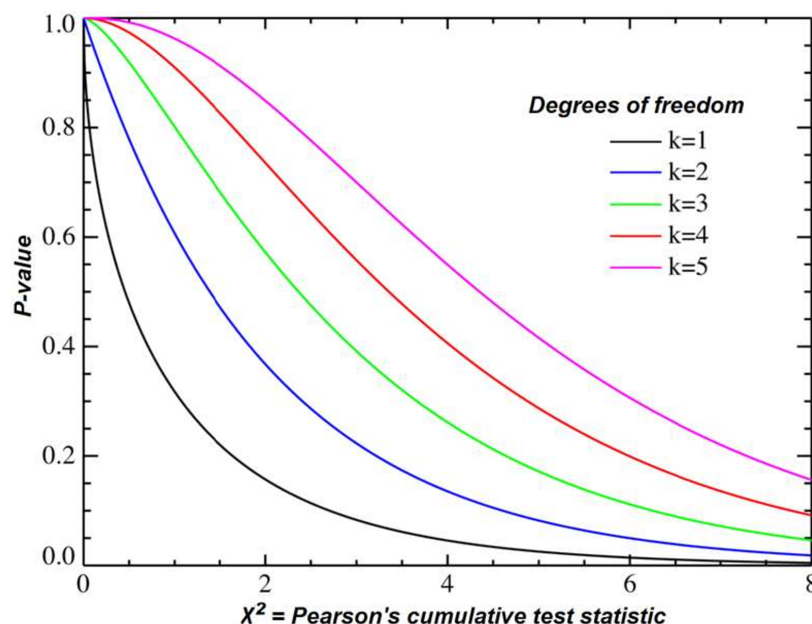


Fig. 10.1 *Chi-Squared Distribution*

In the 19th century, statistical analytical methods were mainly applied in biological data analysis and it was customary for researchers to assume that observations followed a normal distribution. Until the end of 19th century, Pearson noticed the existence of significant skewness within some biological observations. In order to model the observations regardless of being normal or skewed, Pearson formulated the Pearson distribution, a family of continuous probability distributions, which includes the normal distribution and many skewed distributions, and proposed a method of statistical analysis consisting of using the Pearson distribution to model the observation and performing the test of goodness of fit to determine how well the model and the observation really fit.

The chi-squared distribution is continuous probability distribution whose shape is defined by the number of degrees of freedom. It is a right-skew distribution,

but as the number of degrees of freedom increases it approximates the Normal distribution, as shown in Figure 10.2. The chi-squared distribution is important for its use in chi-squared tests. These are often used to test deviations between observed and expected frequencies, or to determine the independence between categorical variables. When conducting a chi-squared test, the probability values derived from chi-squared distributions can be looked up in a statistical table. Figure 1.9 illustrates the chi-squared distribution for various degrees of freedom (df). The distribution becomes less right-skew as the number of degrees of freedom increases.

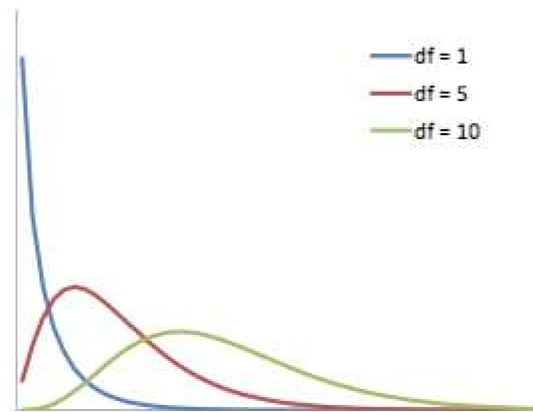


Fig. 10.2 *Chi-Squared Distribution for Various Degrees of Freedom*

There are two types of chi-square tests. Both use the chi-square statistic and distribution for different purposes:

1. A chi-square goodness of fit test determines if a sample data matches a population.
2. A chi-square test for independence compares two variables in a contingency table to see if they are related. In a more general sense, it tests to see whether distributions of categorical variables differ from each another.
   - A very small chi square test statistic means that your observed data fits your expected data extremely well. In other words, there is a relationship.
   - A very large chi square test statistic means that the data does not fit very well. In other words, there is not a relationship.

A chi-square test will give a p-value. The p-value defines that the test results are significant or not. In order to perform a chi-square test and get the p-value, the following information is required:

1. Degrees of freedom. That is just the number of categories minus 1.
2. The alpha level (α). This is chosen by the experimenter/researcher. The usual alpha level is 0.05 (5%), but you could also have other levels like 0.01 or 0.10.

NOTES

10.3 '*t*' TEST

The *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis. The *t*-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. 'Student' was his pen name, so this test is also termed as 'Student *t* Test'.

A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t* distribution. The *t*-test can be used, for example, to determine if the means of two sets of data are significantly different from each other.

Among the most frequently used *t*-tests are:

- A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
- A two-sample location test of the null hypothesis such that the means of two populations are equal. All such tests are usually called Student's *t*-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal, the form of the test used when this assumption is dropped is sometimes called Welch's *t*-test. These tests are often referred to as 'unpaired' or 'independent samples' *t*-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.

The test statistics have the form $t = Z/s$, where Z and s are functions of the data.

Z may be sensitive to the alternative hypothesis, i.e., its magnitude tends to be larger when the alternative hypothesis is true, whereas s is a scaling parameter that allows the distribution of t to be determined.

As an example, in the one-sample *t*-test,

$$t = \frac{Z}{s} = \frac{\overline{X} - \mu}{\hat{\sigma}/\sqrt{n}}$$

Where,

- X is the sample mean from a sample $X_1, X_2, \ldots, X_n$, of size n.
- s is the standard error of the mean.
- $\hat{\sigma}$ is the estimate of the standard deviation of the population.
- μ is the population mean.

The assumptions underlying a *t*-test in its simplest form are that:

- X follows a normal distribution with mean μ and variance σ^2/n.
- s^2 follows a χ^2 distribution with $n - 1$ degrees of freedom.
- Z and s are independent.

t-Distribution

Student's *t*-distribution is a continuous probability distribution with a similar shape to the Normal distribution but with wider tails. The *t*-distributions are used to describe samples which have been drawn from a population, and the exact shape of the distribution varies with the sample size. The smaller the sample size, the more spread out the tails, and the larger the sample size, the closer the *t*-distribution is to the Normal distribution, as shown in Figure 10.3. Whilst in general the Normal distribution is used as an approximation when estimating means of samples from a Normally distribution population, when the same size is small (say $n < 30$), then the *t*-distribution should be used in preference. Figure 10.3 illustrates the *t*-distribution for various sample sizes. As the sample size increases, then the *t*-distribution more closely approximates the Normal.

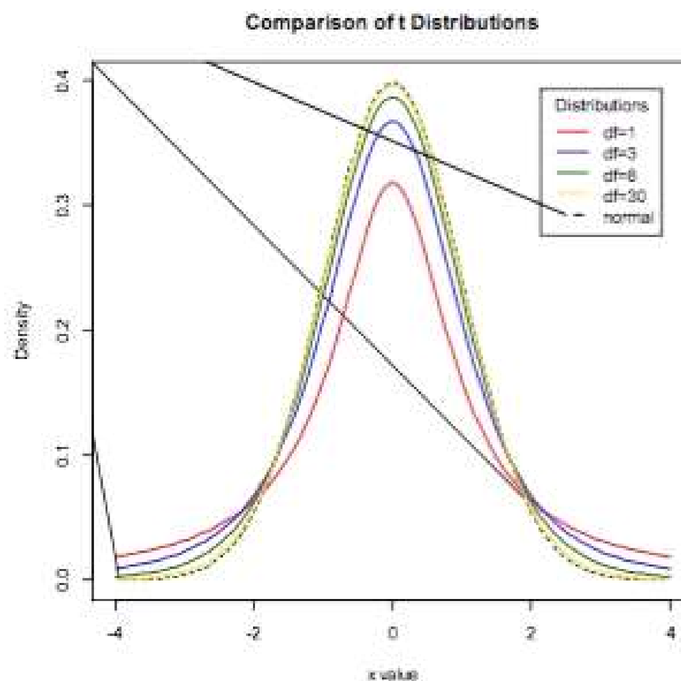


Fig. 10.3 *Comparison of the t-Distribution for Various Sample Sizes*

10.4 ANALYSIS OF VNRIANCE (ANOVA)

ANalysis Of VNriance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA

was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalizes the *t*-test to more than two groups. ANOVA is useful for comparing/testing three or more group means for statistical significance.

The ANOVA is a parametric statistical technique used to compare datasets. This technique was invented by R. A. Fisher, and is thus often also referred to as Fisher's ANOVA. It is similar in application to techniques, such as *t*-test and z-test, in that it is used to compare means and the relative variance between them. However, ANOVA is best applied where more than 2 populations or samples are meant to be compared.

The use of the ANOVA parametric statistical technique involves certain key assumptions, including the following:

1. **Independence of Case:** Independence of case assumption means that the case of the dependent variable should be independent or the sample should be selected randomly. There should not be any pattern in the selection of the sample.
2. **Normality:** Distribution of each group should be normal. The Kolmogorov-Smirnov or the Shapiro-Wilk test may be used to confirm normality of the group.
3. **Homogeneity:** Homogeneity means variance between the groups should be the same. Levene's test is used to test the homogeneity between groups.

If particular data follows the above assumptions, then the ANOVA is the best technique to compare the means of two, or more, populations.

The analysis of ANOVA has following three types:

- **One Way Analysis:** When we are comparing more than three groups based on one factor variable, then it said to be One Way ANOVA. For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.
- **Two Way Analysis:** When factor variables are more than two, then it is said to be Two Way ANOVA. For example, based on working condition and working hours, we can compare whether or not the mean output of three workers is the same.
- **K-Way Analysis:** When factor variables are **K**, then it is said to be the K-Way ANOVA.

The ANOVA is computed using the following key concepts:

- **Sum of Square between Groups:** For the sum of the square between groups, we calculate the individual means of the group, then we take the deviation from the individual mean for each group. And finally, we will take the sum of all groups after the square of the individual group.
- **Sum of Squares within Group:** In order to get the sum of squares within a group, we calculate the grand mean for all groups and then take the deviation from the individual group. The sum of all groups will be done after the square of the deviation.
- **F –Ratio:** To calculate the F-ratio, the sum of the squares between groups will be divided by the sum of the square within a group.
- **Degree of Freedom:** To calculate the degree of freedom between the sums of the squares group, we will subtract one from the number of groups. The sum of the square within the group's degree of freedom will be calculated by subtracting the number of groups from the total observation.
- **BSS df = (g-1)** for BSS is Between the Sum of Squares, where **g** is the group, and **df** is the degree of freedom.
- **WSS df = (N-g)** for WSS is Within the Sum of Squares, where **N** is the total sample size, and **df** is the degree of freedom.
- **Significance:** At a predetermine level of significance (usually at 5%), we will compare and calculate the value with the critical table value. Today, however, computers can automatically calculate the probability value for F-ratio suing the statistical software.

Check Your Progress

1. What is chi-squared test?
2. Explain the *t*-test.
3. Give the uses of *t*-test.
4. What is ANOVA?
5. Elaborate on the types of ANOVA.

10.5 CORRELATION AND REGRESSION ANALYSIS

Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect. Fundamentally, the correlation is a bivariate analysis that measures the

strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and –1. A value of ± 1 indicates a perfect degree of association between the two variables. As the correlation coefficient value goes towards 0, the relationship between the two variables will be weaker. The direction of the relationship is indicated by the sign of the coefficient. A '+' sign indicates a positive relationship and a '–' sign indicates a negative relationship. Correlation is, therefore, a statistical technique that can show whether and how strongly pairs of variables are related, for example, height and weight of an individual, fatty and skinny individual, taller and shorter people, etc. The relationship can be correlated as, people of the same height vary in weight, after analysis you can find that which two people of the population with shorter height is heavier than the taller one. Correlation can define that how much of the variation in peoples' weights is related to their heights.

A perfect positive correlation means that the correlation coefficient is exactly 1 while a perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no relationship at all.

In correlation analysis, a sample correlation coefficient is estimated which is denoted as '*r*'. This ranges between –1 and +1 and quantifies the direction and strength of the linear association between the two variables. The correlation between two variables can be positive, i.e., higher levels of one variable are associated with higher levels of the other or negative, i.e., higher levels of one variable are associated with lower levels of the other.

The sign of the correlation coefficient indicates the direction of the association while the magnitude of the correlation coefficient indicates the strength of the association. For example, a correlation of $r = 0.9$ recommends a strong, positive association between two variables, whereas a correlation of $r = -0.2$ recommends a weak, negative association. A correlation close to zero suggests no linear association between two continuous variables.

The Formula for Correlation is,

$$r = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{\sqrt{\sum (X - \bar{X})^2} \sqrt{(Y - \bar{Y})^2}}$$

Where,

r = Correlation Coefficient

$\bar{X}$ = Average of Observations of Variable X

$\bar{Y}$ = Average of Observations of Variable Y

A correlation between variables indicates that as one variable changes in value, the other variable tends to change in a specific direction. The value of one variable can be used to predict the value of the other variable. For example, height and weight are correlated, hence as height increases the weight also tends to increase.

Scatterplot diagram helps to check for relationships between pairs of continuous data. The scatterplot shown in Figure 10.4 displays the height and weight of teenage girls. Each dot on the graph represents an individual girl and her height and weight on the representative axis.

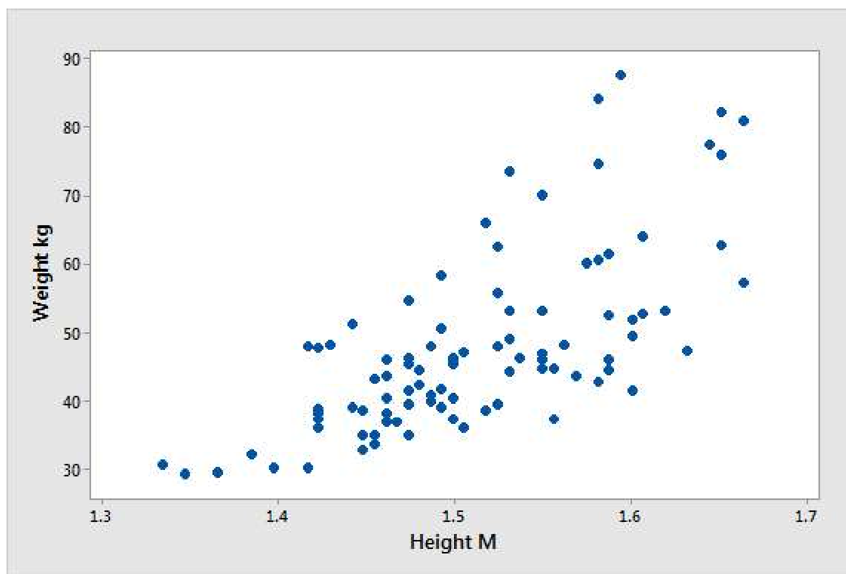


Fig. 10.4 *Scatterplot of the Height and Weight of Teenage Girls*

Figure 10.4 shows that there is a relationship between height and weight. As height increases, the weight may also increase. However, this is not a perfect relationship, for example when a specific height, say 1.5 meters, is considered then there is a range of weights associated with this specific height. However, the common tendency that height and weight increase together is unquestionably observed. Pearson's correlation takes all of the data points on this graph and represents them as a single number.

The following example data in Table 10.1 illustrates the correlation between the Gestational Age and Birth Weight of 17 Infants.

Experimental Example for Studying the Correlation between the Gestational Age and Birth Weight of 17 Infants

An experimental study was conducted on 17 infants to investigate the association or correlation between the gestational age at birth which was measured in weeks

NOTES

and the birth weight which was measured in grams. The observed data was tabulated as shown in Table 10.1.

Table 10.1 *Correlation between the Gestational Age and Birth Weight of 17 Infants*

Infant ID #	Gestational Age (wks)	Birth Weight (gm)
1	34.7	1895
2	36.0	2030
3	29.3	1440
4	40.1	2835
5	35.7	3090
6	42.4	3827
7	40.3	3260
8	37.3	2690
9	40.9	3285
10	38.3	2920
11	38.5	3430
12	41.4	3657
13	39.7	3685
14	39.7	3345
15	41.1	3260
16	38.0	2680
17	38.7	2005

Now we will estimate the association or correlation between the gestational age and infant birth weight from the obtained experimental data. In this example, birth weight is the dependent variable and gestational age is the independent variable. Thus y = Birth Weight and x = Gestational Age. This data is plotted on a graph and is shown in a scatter diagram form as illustrated in the Figure 10.5.

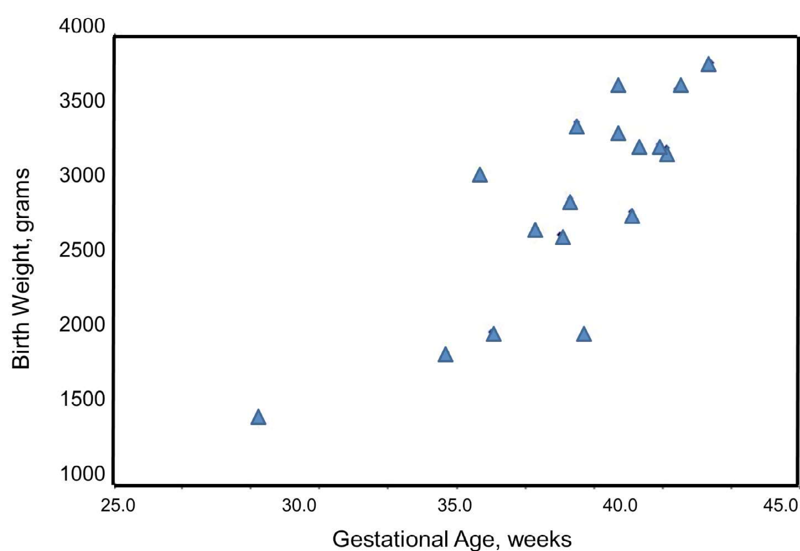


Fig. 10.5 *Scatter Diagram for Gestational Age and Birth Weight of 17 Infants*

NOTES

In the scatter diagram shown in Figure 10.5, each point represents an (x, y) pair, i.e., the gestational age measured in weeks and the birth weight measured in grams. The independent variable 'gestational age' is on the horizontal axis (or X-axis) and the dependent variable 'birth weight' is on the vertical axis (or Y-axis). The scatter plot displays a positive or direct association/correlation between the gestational age and the birth weight. The probability is that the infants with shorter gestational ages are more likely to be born with lower weights while the infants with longer gestational ages are more likely to be born with higher weights.

The formula for the sample correlation coefficient is,

$$r = \frac{\text{Cov}(x, y)}{\sqrt{s_x^2 * s_y^2}}$$

Where Cov(x,y) is the covariance of x and y defined as,

$$\text{Cov}(x, y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}$$

s_x^2 and s_y^2 are the sample variances of x and y, defined as

$$s_x^2 = \frac{\sum (X - \bar{X})^2}{n-1} \text{ and } s_y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1}$$

The variances of x and y measure the variability of the x scores and y scores around their respective sample means ($\bar{X}$ and $\bar{Y}$, considered separately). The covariance measures the variability of the (x, y) pairs around the mean of x and mean of y, considered simultaneously.

To calculate the sample correlation coefficient, we need to compute the variance of gestational age, the variance of birth weight and also the covariance of gestational age and birth weight.

We first summarize the gestational age data as shown below. The mean gestational age is calculated as:

$$\bar{X} = \frac{\sum X}{n} = \frac{652.1}{17} = 38.4$$

To calculate the variance of gestational age, we need to sum the squared deviations (or differences) between each observed gestational age and the mean gestational age. The calculations are summarized in Table 10.2.

NOTES

Table 10.2 *Gestational Age and the Mean Gestational Age*

Infant ID #	Gestational Age	$(X-\bar{X})$	$(X-\bar{X})^2$
1	34.7	-3.7	13.69
2	36.0	-2.4	5.76
3	29.3	-9.1	82.81
4	40.1	1.7	2.89
5	35.7	-2.7	7.29
6	42.4	4.0	16.00
7	40.3	1.9	3.61
8	37.3	-1.1	1.21
9	40.9	2.5	6.25
10	38.3	-0.1	0.01
11	38.5	0.1	0.01
12	41.4	3.0	9.00
13	39.7	1.3	1.69
14	39.7	1.3	1.69
15	41.1	2.7	7.29
16	38.0	-0.4	0.16
17	38.7	0.3	0.09
	$\Sigma X = 652.1$	$\Sigma (X-\bar{X}) = 0$	$\Sigma(X-\bar{X})^2 = 159.45$

The variance of gestational age is given as:

$$s_x^2 = \frac{\sum (X-\bar{X})}{n-1} = \frac{159.45}{16} = 10.0$$

Subsequently, we summarize the birth weight data as follows. The mean birth weight is calculated as:

$$\bar{Y} = \frac{\sum Y}{n} = \frac{49,334}{17} = 290.2$$

The variance of birth weight is calculated in the similar method as we have done for the gestational age. The calculation of birth weight and the mean birth weight is shown in Table 10.3.

Table 10.3 *Birth Weight and the Mean Birth Weight*

Infant ID #	Birth Weight	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$
1	1895	-1007	1,014,049
2	2030	-872	760,384
3	1440	-1462	2,137,444
4	2835	-67	4,489
5	3090	188	35,344
6	3827	925	855,625
7	3260	358	128,164
8	2690	-212	44,944
9	3285	383	146,689
10	2920	18	324
11	3430	528	278,784
12	3657	755	570,025
13	3685	783	613,089
14	3345	443	196,249
15	3260	358	128,164
16	2680	-222	49,284
17	2005	-897	804,609
	$\Sigma Y = 49,334$	$\Sigma (Y-\bar{Y}) = 0$	$\Sigma(Y-\bar{Y})^2 = 7,767,660$

The variance of birth weight is given as:

$$s_y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

NOTES

The covariance is calculated as follows,

$$\text{Cov}(x, y) = \frac{\sum (X - \bar{X})(Y - \bar{Y})}{n-1}$$

To calculate the covariance of gestational age and birth weight, multiply the deviation from the mean gestational age by the deviation from the mean birth weight for each participant (i.e., $(X - \bar{X})(Y - \bar{Y})$).

The calculations are summarized in Table 10.4. In this Table 10.4, we have copied the deviations from the mean gestational age and mean birth weight from the two Tables 10.2 and 10.3 and then multiplied.

Table 10.4 *Covariance of Mean Gestational Age and Mean Birth Weight*

Infant Identification Number	$(X - \bar{X})$	$(Y - \bar{Y})$	$(X - \bar{X})(Y - \bar{Y})$
1	-3.7	-1007	3725.9
2	-2.4	-872	2092.8
3	-9.1	-1462	13,304.2
4	1.7	-67	-113.9
5	-2.7	188	-507.6
6	4.0	925	3700.0
7	1.9	358	680.2
8	-1.1	-212	233.2
9	2.5	383	957.5
10	-0.1	18	-1.8
11	0.1	528	52.8
12	3.0	755	2265.0
13	1.3	783	1017.9
14	1.3	443	575.9
15	2.7	358	966.6
16	-0.4	-222	88.8
17	0.3	-897	-269.1
			$\sum (X - \bar{X})(Y - \bar{Y})$= 28,768.4

The covariance of gestational age and birth weight is given as:

$$s_y^2 = \frac{\sum (Y - \bar{Y})^2}{n-1} = \frac{7,767,660}{16} = 485,578.8.$$

We now calculate the sample correlation coefficient as follows:

$$r = \frac{\text{Cov}(x, y)}{\sqrt{s_x^2 * s_y^2}} = \frac{1798.0}{\sqrt{10.0 * 485,578.8}} = \frac{1798.0}{2199.4} = 0.82.$$

NOTES

The sample correlation coefficient specifies a strong positive correlation.

The sample correlation coefficients range from –1 to +1. In fact, meaningful correlations (i.e., correlations that are clinically or practically significant) can be as small as 0.4 (or –0.4) for positive (or negative) associations. There are also other statistical tests which help to determine whether an observed correlation is statistically significant or not (i.e., statistically significantly different from zero).

Correlation Coefficient

A **correlation coefficient** is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a **sample**, or two components of a multivariate random variable with a known **distribution**.

Several types of correlation coefficient exist, each with their own definition and own range of usability and characteristics. They all assume values in the range from –1 to +1, where ±1 indicates the strongest possible agreement and 0 the strongest possible disagreement.

The correlation coefficient is a statistical measure that calculates the strength of the relationship between the relative movements of two variables. The values range between -1.0 and 1.0. A calculated number greater than 1.0 or less than –1.0 means that there was an error in the correlation measurement. A correlation of –1.0 shows a *perfect negative correlation*, while a correlation of 1.0 shows a *perfect positive correlation*. A correlation of 0.0 shows *no relationship* between the movements of the two variables. The correlation coefficient, r, can be calculated as shown below.

Correlation Coefficient, *r*

The quantity r, called the *linear correlation coefficient*, measures the strength and the direction of a linear relationship between two variables. The linear correlation coefficient is sometimes referred to as the *Pearson product moment correlation coefficient* in honour of its developer Karl Pearson.

The mathematical formula for computing $\boldsymbol{r}$ is:

$$r = \frac{n\sum xy - \left(\sum x\right)\left(\sum y\right)}{\sqrt{n\left(\sum x^2\right) - \left(\sum x\right)^2}\ \sqrt{n\left(\sum y^2\right) - \left(\sum y\right)^2}}$$

Where n is the number of pairs of data.

The value of r is such that $-1 \leq r \leq +1$. The '+' and '–' signs are used for positive linear correlations and negative linear correlations, respectively.

Positive Correlation: If x and y have a strong positive linear correlation, r is close to +1. An r value of exactly +1 indicates a perfect positive fit. Positive values

indicate a relationship between x and y variables such that as values for x increases, values for y also increase.

Negative Correlation: If x and y have a strong negative linear correlation, r is close to -1. An r value of exactly -1 indicates a perfect negative fit. Negative values indicate a relationship between x and y such that as values for x increase, values for y decrease.

No Correlation: If there is no linear correlation or a weak linear correlation, then r is close to 0. A value near zero means that there is a random, nonlinear relationship between the two variables

Remember that r is a dimensionless quantity, i.e., it does not depend on the units employed.

Perfect Correlation: A perfect correlation of $\pm$ 1 occurs only when the data points all lie exactly on a straight line. If $r = +1$, the slope of this line is positive. If $r = -1$, then the slope of this line is negative.

A correlation greater than 0.8 is generally described as *strong*, whereas a correlation less than 0.5 is generally described as *weak*. These values can vary based upon the 'type' of data being examined. A study utilizing scientific data may require a stronger correlation than a study using social science data.

Regression Analysis

Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables. More specifically, regression analysis helps one understand how the typical value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are held fixed.

Definition: Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

Regression analysis is a set of statistical methods used for the estimation of relationships between a dependent variable and one or more independent variables. It can be utilized to assess the strength of the relationship between variables and for modelling the future relationship between them. The regression analysis includes several variations, such as linear, multiple linear, and nonlinear. The most common models are simple linear and multiple linear. Nonlinear regression analysis is commonly used for more complicated data sets in which the dependent and independent variables show a nonlinear relationship.

Linear Model Assumptions for Regression Analysis

Linear regression analysis is based on following six fundamental assumptions:

1. The dependent and independent variables show a linear relationship between the slope and the intercept.
2. The independent variable is not random.
3. The value of the residual (error) is zero.
4. The value of the residual (error) is constant across all observations.
5. The value of the residual (error) is not correlated across all observations.
6. The residual (error) values follow the normal distribution.

Simple Linear Regression

Simple linear regression is a model that assesses the relationship between a dependent variable and one independent variable. The simple linear model is expressed using the following equation:

$$Y = a + bX + \in$$

Where:

Y = Dependent Variable

X = Independent (Explanatory) Variable

a = Intercept

b = Slope

$\in$ = Residual (Error)

Multiple Linear Regression

Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model. The mathematical representation of multiple linear regression is:

$$Y = a + bX_1 + cX_2 + dX_3 + \in$$

Where:

Y = Dependent Variable

X_1, X_2, X_3 = Independent (Explanatory) Variables

a = Intercept

b, c, d = Slopes

ε = Residual (Error)

Multiple linear regression follows the same conditions as the simple linear model. However, since there are several independent variables in multiple linear analysis, there is another mandatory condition for the model named as Non-Collinearity.

NOTES

Non-Collinearity: Independent variables should show a minimum of correlation with each other. If the independent variables are highly correlated with each other, it will be difficult to assess the true relationships between the dependent and independent variables.

Regression takes a group of random variables, thought to be predicting Y, and tries to find a mathematical relationship between them. This relationship is typically in the form of a **straight line (linear regression)** that best approximates all the individual data points. In multiple regression, the separate variables are differentiated by using numbers with subscripts.

10.6 STATISTICAL PACKAGES

Statistical Package for the Social Sciences or SPSS is a computer program used for statistical analysis. Between 2009 and 2010 the premier vendor for SPSS was named Predictive Analytics Software (PASW) Statistics. It was later acquired by International Business Machines (IBM) and as of January 2010 it became 'SPSS: An IBM Company'. The features of SPSS incorporate modules for statistical data analysis, including descriptive statistics, such as plots, frequencies, charts and lists, as well as sophisticated inferential and multivariate statistical procedures, such as ANalysis of VAriance (ANOVA), factor analysis, cluster analysis and categorical data analysis. In addition to statistical analysis, data management and data documentation are unique features of the base software. The various features of SPSS are easily accessible through pull-down menus or can also be programmed using a proprietary Fourth Generation Language (4GL) command syntax language. SPSS is a modular product hence it requires the Base System module to run.

SPSS Statistics 17.0 which is a comprehensive system for analysing data based on the GUI. SPSS Statistics can obtain data from almost any type of file and use them to produce tabulated reports, charts, plots of distributions and trends, descriptive statistics and complex statistical analyses. In addition, below the menus and dialog boxes, SPSS Statistics uses a command language for data analysis.

The syntax editor in SPSS Statistics 17.0 has been entirely redesigned including features, such as auto completion, colour coding, bookmarks and breakpoints. Analysing data using SPSS Statistics is simple and easy. The Data Editor provides a convenient, spreadsheet like method for creating and editing data files. The Data Editor window opens automatically when the user starts SPSS. The dimensions of the data file are determined by the number of cases and variables. You can enter data in any cell. Variable names can be up to 64 bytes long and the first character must be a letter or one of the characters @, # or $. Subsequent characters can be any combination of letters, numbers, non-punctuation characters and a period (.). The user can also specify the level of measurement as scale (numeric data on an interval or ratio scale), ordinal or nominal as per his/her requirements. Nominal and ordinal data can be either string (alphanumeric) or numeric.

10.6.1 About SPSS

SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use. The SPSS Data Editor is very valuable and is specifically designed for performing statistical tests, such as correlation, regression, *t*-test, hypotheses, chi-square and Analysis of Variance or ANOVA. It also helps a researcher to make useful data entries, find frequency counts, sort and rearrange data, etc.

The SPSS features available with the software package can be accessed with the help of pull-down menus or can be programmed using a licensed 4GL (Fourth Generation Language) command syntax language. The advantage of command syntax programming language is that it helps in data reproducibility, simplifying repetitive tasks, performing complex data manipulations and analyses. In addition, the user can program specific syntax for some complex applications which are not available in the predefined menu structure. The command syntax can also be generated by pull-down menu interface and can be displayed in the output. To syntax can be made visible to the user by changing the default settings. It can also be pasted into a syntax file with the help of 'Paste' button which is available in each menu.

SPSS can read and write data from American Standard Code for Information Interchange (ASCII) text files including hierarchical files, other statistics packages, spreadsheets and databases. SPSS can also be used to read and write to external relational database tables using Open Database Connectivity (ODBC) and Sequential Query Language (SQL). Statistical output is in the licensed file format with the file extension name as **.spv** which supports pivot tables. The output can be exported to Microsoft Word and can be acquired as data, as text, Portable Document Format (PDF), Microsoft Excel Spreadsheet (XLS), Hyper Text Mark-up Language (HTML), Extensible Mark-up Language (XML), SPSS dataset or in the graphic image formats [Joint Photographic Experts Group (JPEG), Portable Network Graphics (PNG), Bitmap (BMP) and Enhanced Meta Files(EMF)].

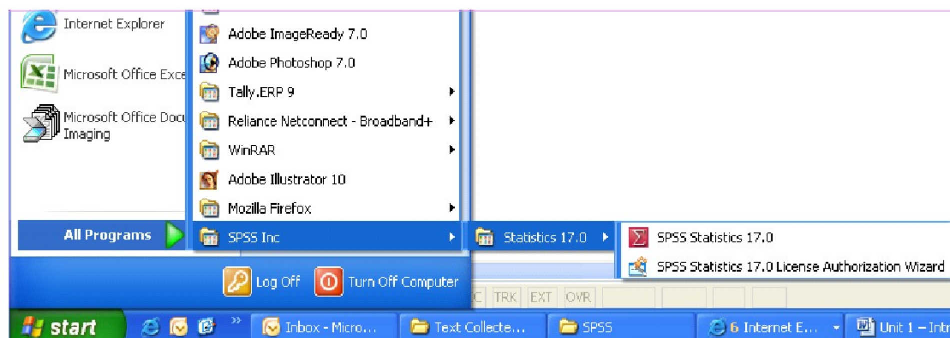
The SPSS is based on Graphical User Interface (GUI) which supports the two data editor views, the Data View and the Variable View. The user can toggle between the two views just by selecting one of the two tabs that appear in the bottom left of the SPSS window and clicking on it. The 'Data View' exhibits a view in the form of a spreadsheet as the cases (rows) and variables (columns). Only two data types can be defined in SPSS Statistics, i.e., the numeric data type and the text or 'String' data type. All data processing processes appears in sequence case-by-case through the file. You can match the files on the basis of one-to-one and one-to-many, but not many-to-many. In SPSS, the data cells simply hold numbers or text. You cannot store the formulas in these cells. The 'Variable View'

NOTES

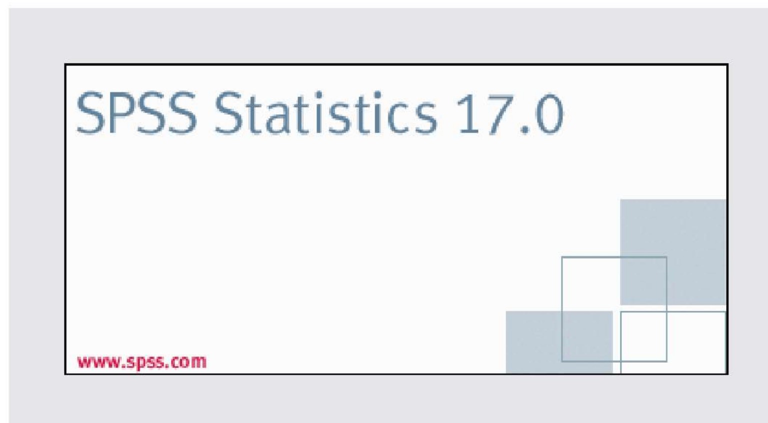
exhibits the metadata dictionary in which each row represents a variable to display the variable name, variable label, value label(s), print width, measurement type and other associated characteristics. In both views, you can manually edit the cells, define file structure and do data entry without using the command syntax for smaller datasets. Large datasets, such as statistical surveys are created using data entry software or entered by scanning using Optical Character Recognition (OCR) and Optical Mark Recognition (OMR) software. Using a 'Macro' language command language subroutines can be written. A Python programmability extension is used to access the information in the data dictionary and dynamically build command syntax programs.

10.6.2 Working with SPSS

Install SPSS on your computer and create a **shortcut** menu on your desktop and to directly start the package by clicking on the SPSS icon. Alternatively, you can go to the **Start → All Programs → SPSS Inc → Statistics 17.0 → SPSS Statistics 17.0** as shown:



When you click on the SPSS Statistics 17.0, the following screen will appear to start SPSS program:

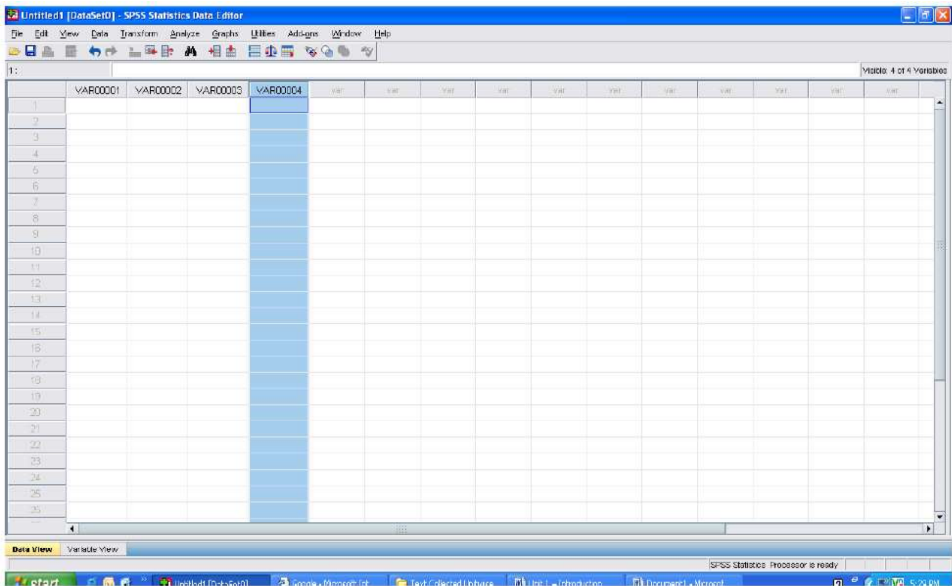


As discussed earlier, the SPSS Data Editor has two main 'Views'. User can enter data in the Data View while in the Variable View, user can select the name, type, maximum number of letters per cell ('Width'), number of decimal points, label, width of cell ('Columns'), alignment within the cell ('Align') and

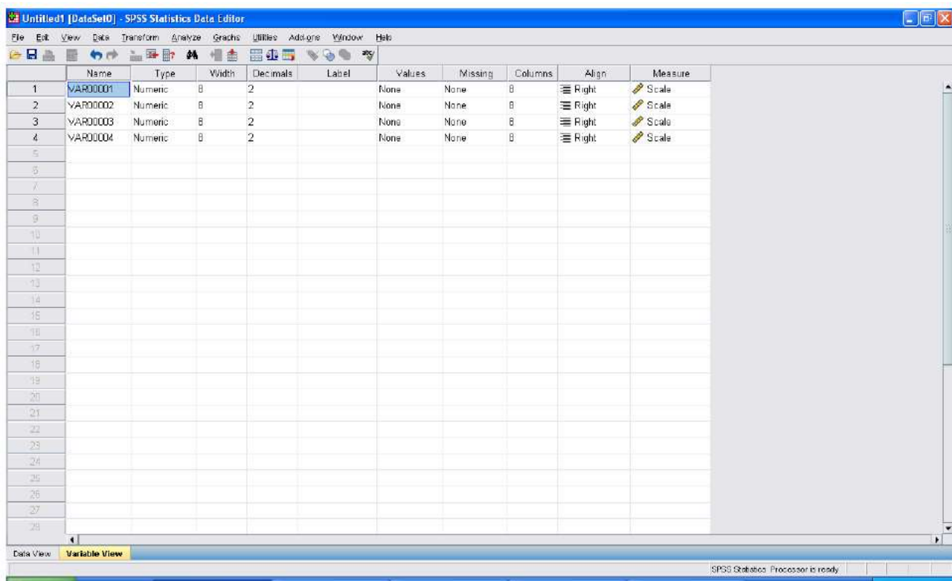
NOTES

whether or not the variable is nominal, ordinal or 'Scale' ('Measure'). The user can also categorize entries into labels (in the 'Values' column) and mark entries as invalid (in the 'Missing' column) in the Variable View. Using SPSS for Windows the user can perform almost any statistics calculation in combination with pointing and clicking on the menus and various specific interactive dialog boxes. Both the SPSS views are shown on the next page.

Data View



Variable View



10.6.3 SPSS Statistics 17.0

SPSS Statistics 17.0 is a comprehensive system for analysing data based on the GUI. SPSS Statistics can acquire data from almost any form of file and use them to create tabulated reports, charts, plots of distributions and trends, descriptive statistics and complex statistical analyses. SPSS Statistics Base 17.0 provides examples in the Help system which is automatically installed with the software. In addition, below the menus and dialog boxes, SPSS Statistics uses a command language for data analysis.

SPSS Statistics has a powerful statistical analysis and data management system in a graphical environment. It also has descriptive menus and simple dialog boxes which help the users to accomplish the task just by pointing and clicking the mouse. In addition to the simple point-and-click interface for statistical analysis, SPSS Statistics provides the following features:

- **Data Editor:** The Data Editor is similar to multipurpose spreadsheet system and is used to define, enter, edit and display data.
- **Viewer:** The Viewer helps to browse the results, show and hide selective outputs, modify the display order results, shift presentation quality tables and charts to and from other applications.
- **Multidimensional Pivot Tables:** The multidimensional pivot tables display the output results in the form which look alive. Users can explore tables by rearranging rows, columns and layers. It is also easy to compare the groups. It is done by splitting the table so that only one group is displayed at a time.
- **High Resolution Graphics:** High resolution, full colour pie charts, bar charts, histograms, scatter plots, 3D graphics, etc., are built-in standard features.
- **Database Access:** The user can directly recover information from databases by using the Database Wizard omitting the complex SQL queries.
- **Data Transformations:** Transformation features help to find the data organized for analysis. You can also subset data and files to combine categories, add, aggregate, merge, split, transpose and much more.
- **Online Help:** A comprehensive abstract of context sensitive Help topics are available in dialog boxes to guide the users while performing specific tasks, pop-up definitions in pivot table results, explaining statistical terms. The Statistics Coach helps the users to find the required procedures whereas Case Studies provide hands-on examples for using statistical procedures and to interpret the results.
- **Command Language:** Most of the tasks in SPSS Statistics are completed with the help of simple point-and-click actions. It also provides

NOTES

a powerful command language which permits the user to save and automate various common tasks. The command language also provides several functionalities which are not available in the menus and dialog boxes. Complete command syntax documentation is incorporated into the overall Help system.

10.6.4 What's New in SPSS Statistics Version 17.0?

The following are the enhanced features available in the new version of SPSS Statistics 17.0:

- **New Syntax Editor:** The syntax editor in SPSS Statistics 17.0 has been entirely redesigned including features such as auto completion, colour coding, bookmarks and breakpoints. Auto completion feature provides the user a list of valid command names, subcommands and keywords hence the user spends less time referring to syntax charts. Colour coding feature permits the user to spot unrecognized terms and some common syntactical errors quickly. Bookmarks feature permits the user to speedily navigate huge command syntax files. Breakpoints feature permits the user to stop execution at specific points for inspecting data or output prior to proceeding.
- **Custom Dialog Builder:** The Custom Dialog Builder permits the user to create and manage custom dialogs for creating command syntax.
- **Multiple Language Support:** In addition to the capability to modify the output language the user can now modify the user interface language.
- **Codebook:** The Codebook method accounts the dictionary information, such as variable names, variable labels, value labels, missing values and summary statistics for all or specific variables and manifold response sets in the active dataset. For nominal and ordinal variables, and manifold response sets, summary statistics include counts and percent's. For scale variables, summary statistics include mean, standard deviation and quartiles.
- **Nearest Neighbour Analysis:** Nearest Neighbour analysis is a technique for categorizing cases based on their similarity to other cases. In machine learning, it was used to identify patterns of data without involving an accurate match to any stored patterns or cases. Similar cases are close to each other and dissimilar cases are isolated from each other. Thus, the distance between two cases is a measure of their dissimilarity.
- **Multiple Imputation:** The Multiple Imputation method executes multiple imputation of missing data values. The dataset having missing values give outputs as one or more datasets in which missing values are substituted with plausible estimates. The pooled results obtained when other procedures are run. This technique also summarizes missing values in the working dataset. This feature is available in the Missing Values add-on option.

NOTES

- **RFM Analysis:** Recency, Frequency, Duration (RFM) analysis is the abbreviated form of Recency, Frequency, and Monetary analysis. This method is used to recognise existing customers who are most probable to respond to a new offer and is frequently used in direct marketing. This feature is available in the EZ RFM add-on option. The fundamental principle of RFM analysis is for customers who have purchased recently, have made more purchases and are more likely to respond to your offering than other customers who have purchased less recently, less often and in smaller amounts. Basically, RFM analysis uses information about customers' past behaviour that is easily tracked and readily available. Recency is how long ago the customer last made a purchase. Frequency is how many purchases the customer has made (sometimes within a specified time period, such as average number of purchases per year). Monetary is total amount spent by the customer (sometimes within a specified time period).
- **Categorical Regression Enhancements:** Categorical Regression has been enhanced and included regularization and resampling techniques for accurately assessing and improving predictions. Jointly, these new methods feasibly create state-of-the-art models even for high volume data where there are more variables than observations. This feature is available in the Categories add-on option.
- **Graph Board:** Graph board are visualizations which include graphs, charts and plots created using a visualization template. SPSS Statistics 17.0 provides built-in new visualisation templates which are effectively custom visualisation types.
- **Exporting Output:** The following output export format options and control over exported contents are available in SPSS Statistics version 17.0:
  - To wrap or shrink wide table in Word documents.
  - To create new worksheets or append data to existing worksheets in an Excel workbook.
  - To save output export specifications in the form of command syntax with the OUTPUT EXPORT command. All the features for exporting output in the Export Output dialog are available in command syntax.
  - The Output Management System (OMS) supports the additional output formats, such as Word, Excel and PDF.
- **Shift Values:** Shift Values generates new variables that hold the values of existing variables from preceding or subsequent cases.
- **Aggregate Enhancements:** This feature allows the user to use the aggregate method without specifying a break variable.
- **Median Function:** A median function is available for computing the median value across selected variables for each case.

Check Your Progress

6. What is correlation?
7. What does a perfect positive correlation means?
8. Define correlation coefficient.
9. Differentiate between simple linear regression and multiple linear regression analysis.
10. Explain about the statistical packages.

NOTES

10.7 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.
2. The *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis. The *t*-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. 'Student's was his pen name, so this test is also termed as 'Student *t* -Test'. A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t*- distribution.
3. Among the most frequently used t-tests are:
   - A one-sample location test of whether the mean of a population has a value specified in a null hypothesis.
   - A two-sample location test of the null hypothesis such that the means of two populations are equal. All such tests are usually called Student's *t*-tests, though strictly speaking that name should only be used if the variances of the two populations are also assumed to be equal, the form of the test used when this assumption is dropped is sometimes called Welch's *t*-test. These tests are often referred to as 'unpaired' or 'independent samples' *t*-tests, as they are typically applied when the statistical units underlying the two samples being compared are non-overlapping.

4. ANalysis Of VAriance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation.
5. The analysis of ANOVA has following three types:
   - One Way Analysis: When we are comparing more than three groups based on one factor variable, then it said to be One Way ANOVA. For example, if we want to compare whether or not the mean output of three workers is the same based on the working hours of the three workers.
   - Two Way Analysis: When factor variables are more than two, then it is said to be Two Way ANOVA. For example, based on working condition and working hours, we can compare whether or not the mean output of three workers is the same.
   - K-Way Analysis: When factor variables are k, then it is said to be the k-Way ANOVA.
6. Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect. Fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.
7. A perfect positive correlation means that the correlation coefficient is exactly 1 while a perfect negative correlation means that two assets move in opposite directions, while a zero correlation implies no relationship at all.
8. A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
9. Simple linear regression is a model that assesses the relationship between a dependent variable and one independent variable.

   Multiple linear regression analysis is essentially similar to the simple linear model, with the exception that multiple independent variables are used in the model.
10. SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on

NOTES

computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies. The statistical package SPSS is used to perform quantitative research in social science because it is easy to use.

10.8 SUMMARY

- A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.
- In the standard applications of χ^2 test, the observations are classified into mutually exclusive classes, and there is some theory, or say null hypothesis, which gives the probability that any observation falls into the corresponding class. The purpose of the test is to evaluate how likely the observations that are made would be, assuming the null hypothesis is true.
- Chi-squared tests are often constructed from a sum of squared errors, or through the sample variance. Test statistics that follow a chi-squared distribution arise from an assumption of independent normally distributed data, which is valid in many cases due to the central limit theorem. A chi-squared test can be used to attempt rejection of the null hypothesis that the data are independent.
- The chi-squared distribution is important for its use in chi-squared tests. These are often used to test deviations between observed and expected frequencies, or to determine the independence between categorical variables.
- The *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis. The *t*-statistic was introduced in 1908 by William Sealy Gosset, a chemist working for the Guinness brewery in Dublin, Ireland. 'Student's was his pen name, so this test is also termed as 'Student *t* -Test'.
- A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known. When the scaling term is unknown and is replaced by an estimate based on the data, the test statistics (under certain conditions) follow a Student's *t*- distribution.
- Student's *t*-distribution is a continuous probability distribution with a similar shape to the Normal distribution but with wider tails. The *t*-distributions are used to describe samples which have been drawn from a population, and the exact shape of the distribution varies with the sample size.

NOTES

- Analysis Of VAriance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'Variation' among and between groups, used to analyse the differences among group means in a sample. ANOVA was developed by statistician and evolutionary biologist Ronald Fisher.
- ANOVA was developed by statistician and evolutionary biologist Ronald Fisher. The ANOVA is based on the law of total variance, where the observed variance in a particular variable is partitioned into components attributable to different sources of variation. In its simplest form, ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalises the t-test to more than two groups.
- ANOVA provides a statistical test of whether the population means of several groups are equal, and therefore generalises the t-test to more than two groups. ANOVA is useful for comparing/testing three or more group means for statistical significance.
- The ANOVA is a parametric statistical technique used to compare datasets. This technique was invented by R. A. Fisher, and is thus often also referred to as Fisher's ANOVA. It is similar in application to techniques, such as *t*-test and *z*-test, in that it is used to compare means and the relative variance between them.
- Correlation is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant rate. It is a common tool for describing simple relationships without making a statement about cause and effect.
- The correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship. In terms of the strength of relationship, the value of the correlation coefficient varies between +1 and –1. A value of ± 1 indicates a perfect degree of association between the two variables.
- A correlation coefficient is a numerical measure of some type of correlation, meaning a statistical relationship between two variables. The variables may be two columns of a given data set of observations, often called a sample, or two components of a multivariate random variable with a known distribution.
- Regression analysis is a set of statistical processes for estimating the relationships among variables. It includes many techniques for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.
- Regression is a statistical measurement used in finance, investing and other disciplines that attempts to determine the strength of the relationship between one dependent variable (usually denoted by Y) and a series of other changing variables (known as independent variables).

NOTES

- SPSS can read and write data from American Standard Code for Information Interchange (ASCII) text files including hierarchical files, other statistics packages, spreadsheets and databases. SPSS can also be used to read and write to external relational database tables using Open Database Connectivity (ODBC) and Sequential Query Language (SQL).
- The features of SPSS incorporate modules for statistical data analysis, including descriptive statistics, such as plots, frequencies, charts and lists, as well as sophisticated inferential and multivariate statistical procedures, such as ANalysis of VAriance (ANOVA), factor analysis, cluster analysis and categorical data analysis.
- SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies.
- The SPSS is based on Graphical User Interface (GUI) which supports the two data editor views, the Data View and the Variable View. The user can toggle between the two views just by selecting one of the two tabs that appear in the bottom left of the SPSS window and clicking on it. The 'Data View' exhibits a view in the form of a spreadsheet as the cases (rows) and variables (columns).

10.9 KEY WORDS

- **Chi-Squared Test:** A Chi-Squared Test, also written as χ^2 test, is any statistical hypothesis test where the sampling distribution of the test statistic is a chi-squared distribution when the null hypothesis is true. Without other prerequisite, the 'chi-squared test' often is used as short for Pearson's chi-squared test. The chi-squared test is used to determine whether there is a significant difference between the expected frequencies and the observed frequencies in one or more groups/categories.
- ***t*-test:** The *t*-test is any statistical hypothesis test in which the test statistic follows a Student's *t*-distribution under the null hypothesis. A *t*-test is most commonly applied when the test statistic would follow a normal distribution if the value of a scaling term in the test statistic were known.
- **ANOVA:** ANalysis Of VAriance (ANOVA) is a collection of statistical models and their associated estimation procedures, such as the 'Variation' among and between groups, used to analyse the differences among group means in a sample.
- **Correlation:** It is a statistical measure that expresses the extent to which two variables are linearly related, i.e., they change together at a constant

rate, fundamentally, the correlation is a bivariate analysis that measures the strength of association between two variables and the direction of the relationship.

- **Regression analysis:** It is a set of statistical processes for estimating the relationships among variables, it is used for modelling and analysing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables.
- **SPSS:** SPSS is abbreviated term for Statistical Package for the Social Sciences and is used for data management and analysis. This program is used on computers for statistical analysis in social science by government, market researchers, education researchers, health researchers and survey companies.

NOTES

10.10 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain about the Chi-Square.
2. Give the uses of Chi-square test.
3. What is *t*-test?
4. Interpret the *t*-distribution.
5. When the *t*-test and Chi-square tests are used?
6. Define the term ANOVA.
7. Why is ANOVA used for statistical analysis?
8. What does the sign of the correlation coefficient indicates?
9. Define correlation in terms of the strength of relationship.
10. What is perfect correlation?
11. Elaborate on the regression analysis.
12. Comprehend the simple linear regression analysis.
13. Explain about the statistical packages.

Long-Answer Questions

1. Explain in detail about the Chi-square test.
2. Analyse the *t*-test and *t*-distribution.
3. Briefly discuss about ANalysis Of VAriance giving its uses.
4. Describe the correlation and regression analysis with appropriate examples.
5. Discuss about the statistical packages and SPSS statistics version 17.0.

NOTES

10.11 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

BLOCK - IV
BIOINFORMATICS

NOTES

UNIT 11 INTRODUCTION TO BIOINFORMATICS

Structure

11.0 INTRODUCTION

Bioinformatics is an interdisciplinary field mainly involving molecular biology and genetics, computer science, mathematics, and statistics. Fundamentally, 'Bioinformatics' is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data.

Bioinformatics has become an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques, such as image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in comparing, analysing and interpreting genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyse and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, proteins as well as biomolecular interactions.

In this unit, you will study about the introduction to bioinformatics, medical, informatics, cheminformatics and pharmacoinformatics.

11.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the significance of bioinformatics
- Understand the importance of bioinformatics in medical science
- Know about the informatics, cheminformatics and pharmacoinformatics

11.2 BIOINFORMATICS: BASIC CONCEPTS

Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Common uses of bioinformatics include the identification of genes and Single Nucleotide Polymorphisms (SNPs).

Bioinformatics has become an important part of many areas of biology. In experimental molecular biology, bioinformatics techniques, such as image and signal processing allow extraction of useful results from large amounts of raw data. In the field of genetics, it aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation. Bioinformatics tools aid in comparing, analysing and interpreting genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology. At a more integrative level, it helps analyse and catalogue the biological pathways and networks that are an important part of systems biology. In structural biology, it aids in the simulation and modeling of DNA, RNA, proteins as well as biomolecular interactions.

Historically, the term 'Bioinformatics' did not mean what it means today. Paulien Hogeweg and Ben Hesper coined it in 1970 to refer to the study of information processes in biotic systems. This definition placed bioinformatics as a field parallel to biochemistry (the study of chemical processes in biological systems).

Sequences

Computers became essential in molecular biology when protein sequences became available after Frederick Sanger determined the sequence of insulin in the early

NOTES

1950s. Comparing multiple sequences manually was not possible. A pioneer in the field was Margaret Oakley Dayhoff. She compiled one of the first protein sequence databases, initially published as books and pioneered methods of sequence alignment and molecular evolution. Another early contributor to bioinformatics was Elvin A. Kabat, who pioneered biological sequence analysis in 1970 with his comprehensive volumes of antibody sequences released with Tai Te Wu between 1980 and 1991. In the 1970s, new techniques for sequencing DNA were applied to bacteriophage MS2 and øX174, and the extended nucleotide sequences were then parsed with informational and statistical algorithms. These studies illustrated that well known features, such as the coding segments and the triplet code, are revealed in straightforward statistical analyses and were thus proof of the concept that bioinformatics would be perceptive.

Figure 11.1 illustrates the sequences of genetic material that are frequently used in bioinformatics and are easier to manage using computers than manually.

5'ATGACGTGGGGA3'
3'TACTGCACCCCT5'

Fig. 11.1 *Sequences of Genetic Material*

The field of bioinformatics includes the analysis and interpretation of various types of data. This also includes nucleotide and amino acid sequences, protein domains, and protein structures. The actual process of analysing and interpreting data is referred to as computational biology.

Following are the important sub-disciplines within bioinformatics and computational biology:

- Development and implementation of computer programs that enable efficient access to, management and use of, various types of information.
- Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.

The primary goal of bioinformatics is to enhance the understanding of biological processes. Examples include pattern recognition, data mining, machine learning algorithms, visualization, sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.

NOTES

Relation of Bioinformatics to Other Fields

Bioinformatics is a science field that is similar to but distinct from biological computation, while it is often considered synonymous to computational biology. Biological computation uses bioengineering and biology to build biological computers, whereas bioinformatics uses computation to better understand biology. Bioinformatics and computational biology involve the analysis of biological data, particularly DNA, RNA, and protein sequences. The field of bioinformatics experienced explosive growth starting in the mid-1990s, driven largely by the Human Genome Project and by rapid advances in DNA sequencing technology.

Analysing biological data to produce meaningful information involves writing and running software programs that use algorithms from graph theory, artificial intelligence, soft computing, data mining, image processing, and computer simulation. The algorithms in turn depend on theoretical foundations such as discrete mathematics, control theory, system theory, information theory, and statistics.

Sequence Analysis

Since the Phage Φ-X174 was sequenced in 1977, the DNA sequences of thousands of organisms have been decoded and stored in databases. This sequence information is analysed to determine genes that encode proteins, RNA genes, regulatory sequences, structural motifs, and repetitive sequences. A comparison of genes within a species or between different species can show similarities between protein functions, or relations between species (the use of molecular systematics to construct phylogenetic trees). With the growing amount of data, it long ago became impractical to analyse DNA sequences manually. Computer programs, such as BLAST are used regularly and characteristically to search sequences—as of 2008, from more than 260,000 organisms, containing over 190 billion nucleotides.

DNA Sequencing

Before sequences can be analysed they have to be obtained from the data storage bank. DNA sequencing is still a non-trivial problem as the raw data may be noisy or afflicted by weak signals. Algorithms have been developed for base calling for the various experimental approaches to DNA sequencing.

Sequence Assembly

Most DNA sequencing techniques produce short fragments of sequence that need to be assembled to obtain complete gene or genome sequences. The so-called shotgun sequencing technique (which was used, for example, by The Institute for Genomic Research (TIGR) to sequence the first bacterial genome, ***Haemophilus influenzae***) generates the sequences of many thousands of small DNA fragments (ranging from 35 to 900 nucleotides long, depending on the sequencing technology). The ends of these fragments overlap and, when aligned properly by a genome assembly program, can be used to reconstruct the complete genome. Shotgun sequencing yields sequence data quickly, but the task of assembling the fragments can be quite complicated for larger genomes. For a genome as large as the human

genome, it may take many days of CPU time on large-memory, multiprocessor computers to assemble the fragments, and the resulting assembly usually contains numerous gaps that must be filled in later.

NOTES

Genome Annotation

In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of sequencing has ceased to pose a bottleneck. Annotation is made possible by the fact that genes have recognisable start and stop regions, although the exact sequence found in these regions can vary between genes.

The first description of a comprehensive genome annotation system was published in 1995 by the team at The Institute for Genomic Research that performed the first complete sequencing and analysis of the genome of a free-living organism, the bacterium ***Haemophilus influenzae***. Owen White designed and built a software system to identify the genes encoding all proteins, transfer RNAs, ribosomal RNAs (and other sites) and to make initial functional assignments. Most current genome annotation systems work similarly, but the programs available for analysis of genomic DNA, such as the GeneMark program trained and used to find protein-coding genes in ***Haemophilus influenzae***, are constantly changing and improving.

Computational Evolutionary Biology

Evolutionary biology is the study of the origin and descent of species, as well as their change over time. Informatics has assisted evolutionary biologists by enabling researchers to:

- Trace the evolution of a large number of organisms by measuring changes in their DNA, rather than through physical taxonomy or physiological observations alone.
- Compare entire genomes, which permits the study of more complex evolutionary events, such as gene duplication, horizontal gene transfer, and the prediction of factors important in bacterial speciation.
- Build complex computational population genetics models to predict the outcome of the system over time.
- Track and share information on an increasingly large number of species and organisms.

Comparative Genomics

The core of comparative genome analysis is the establishment of the correspondence between genes (Orthology Analysis) or other genomic features in different organisms. It is these intergenomic maps that make it possible to trace the evolutionary processes responsible for the divergence of two genomes. A multitude of

NOTES

evolutionary events acting at various organizational levels shape genome evolution. At the lowest level, point mutations affect individual nucleotides. At a higher level, large chromosomal segments undergo duplication, lateral transfer, inversion, transposition, deletion and insertion. Ultimately, whole genomes are involved in processes of hybridization, polyploidization and endosymbiosis, often leading to rapid speciation. The complexity of genome evolution poses many exciting challenges to developers of mathematical models and algorithms, who have recourse to a spectrum of algorithmic, statistical and mathematical techniques, ranging from exact, heuristics, fixed parameter and approximation algorithms for problems based on parsimony models to Markov chain Monte Carlo algorithms for Bayesian analysis of problems based on probabilistic models. Many of these studies are based on the detection of sequence homology to assign sequences to protein families.

Genetics of Disease

With the advent of next-generation sequencing we are obtaining enough sequence data to map the genes of complex diseases infertility, breast cancer or Alzheimer's disease. Genome-wide association studies are a useful approach to pinpoint the mutations responsible for such complex diseases. Through these studies, thousands of DNA variants have been identified that are associated with similar diseases and traits. Furthermore, the possibility for genes to be used at prognosis, diagnosis or treatment is one of the most essential applications. Many studies are discussing both the promising ways to choose the genes to be used and the problems and pitfalls of using genes to predict disease presence or prognosis.

Analysis of Mutations in Cancer

In cancer, the genomes of affected cells are rearranged in complex or even unpredictable ways. Massive sequencing efforts are used to identify previously unknown point mutations in a variety of genes in cancer. Bioinformaticians endure to produce specialized automated systems to manage the sheer volume of sequence data produced, and they create new algorithms and software to compare the sequencing results to the growing collection of human genome sequences and germline polymorphisms. New physical detection technologies are employed, such as oligonucleotide microarrays to identify chromosomal gains and losses (called comparative genomic hybridization), and single-nucleotide polymorphism arrays to detect known point mutations. These detection methods simultaneously measure several hundred thousand sites throughout the genome, and when used in high-throughput to measure thousands of samples, generate terabytes of data per experiment. Again the massive amounts and new types of data generate new opportunities for bioinformaticians. The data is often found to contain considerable variability, or noise, and thus **Hidden Markov model** and **change-point analysis** methods are being developed to infer real copy number changes.

Two important principles can be used in the analysis of cancer genomes bioinformatically pertaining to the identification of mutations in the exome. First,

cancer is a disease of accumulated somatic mutations in genes. Second cancer contains driver mutations which need to be distinguished from passengers.

The next-generation sequencing technology in bioinformatics, study how the cancer genomics could drastically change, and to create a more flexible process for classifying types of cancer by analysis of cancer driven mutations in the genome. Furthermore, tracking of patients while the disease progresses may be possible in the future with the sequence of cancer samples. Another type of data that requires novel informatics development is the analysis of lesions found to be recurrent among many tumors.

NOTES

Gene and Protein Expression

The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, Expressed cDNA Sequence Tag (EST) sequencing, Serial Analysis of Gene Expression (SAGE) tag sequencing, Massively Parallel Signature Sequencing (MPSS), RNA-Seq, also known as 'Whole Transcriptome Shotgun Sequencing (WTSS)' or various applications of multiplexed in situ hybridization. All of these techniques are extremely noise-prone and/or subject to bias in the biological measurement, and a major research area in computational biology involves developing statistical tools to separate signal from noise in high-throughput gene expression studies. Such studies are often used to determine the genes implicated in a disorder: one might compare microarray data from cancerous epithelial cells to data from non-cancerous cells to determine the transcripts that are up-regulated and down-regulated in a particular population of cancer cells.

Analysis of Protein Expression

Protein microarrays and High Throughput (HT) Mass Spectrometry (MS) can provide a snapshot of the proteins present in a biological sample. Bioinformatics is very much involved in making sense of protein microarray and HT MS data; the former approach faces similar problems as with microarrays targeted at mRNA, the latter involves the problem of matching large amounts of mass data against predicted masses from protein sequence databases, and the complicated statistical analysis of samples where multiple, but incomplete peptides from each protein are detected. Cellular protein localization in a tissue context can be achieved through affinity proteomics displayed as spatial data based on immunohistochemistry and tissue microarrays.

Analysis of Regulation

Gene regulation is the complex orchestration of events by which a signal, potentially an extracellular signal such as a hormone, eventually leads to an increase or decrease in the activity of one or more proteins. Bioinformatics techniques have been applied to explore various steps in this process. For example, gene expression can be regulated by nearby elements in the genome. Promoter analysis involves the identification and study of sequence motifs in the DNA surrounding the coding

NOTES

region of a gene. These motifs influence the extent to which that region is transcribed into mRNA. Enhancer elements far away from the promoter can also regulate gene expression, through three-dimensional looping interactions. These interactions can be determined by bioinformatic analysis of chromosome conformation capture experiments.

Expression data can be used to infer gene regulation, one might compare microarray data from a wide variety of states of an organism to form hypotheses about the genes involved in each state. In a single cell organism, one might compare stages of the cell cycle, along with various stress conditions (heat shock, starvation, etc.). The clustering algorithms can be applied to that expression data to determine which genes are co-expressed. For example, the upstream regions (promoters) of co-expressed genes can be searched for over-represented regulatory elements. Examples of clustering algorithms applied in gene clustering are k-means clustering, Self-Organizing Maps (SOMs), hierarchical clustering, and consensus clustering methods.

Analysis of Cellular Organization

Several approaches have been developed to analyse the location of organelles, genes, proteins, and other components within cells. This is relevant as the location of these components affects the events within a cell and thus helps us to predict the behaviour of biological systems. A gene ontology category, cellular component, has been devised to capture subcellular localization in many biological databases.

Microscopy and Image Analysis

Microscopic pictures allow us to locate both organelles as well as molecules. It may also help us to distinguish between normal and abnormal cells, for example in cancer.

Protein Localization

The localization of proteins helps us to evaluate the role of a protein. For instance, if a protein is found in the nucleus it may be involved in gene regulation or splicing. By contrast, if a protein is found in mitochondria, it may be involved in respiration or other metabolic processes. Protein localization is thus an important component of protein function prediction. There are well developed protein subcellular localization prediction resources available, including protein subcellular location databases, and prediction tools.

Nuclear Organization of Chromatin

Data from high throughput chromosome conformation capture experiments, such as Hi-C (experiment) and ChIA-PET, can provide information on the spatial proximity of DNA loci. Analysis of these experiments can determine the three-dimensional structure and nuclear organization of chromatin. Bioinformatic challenges in this field include partitioning the genome into domains, such as Topologically Associating Domains (TADs) that are organised together in three-dimensional space.

NOTES

Structural Bioinformatics

Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. Evidently, there are exceptions, such as the bovine spongiform encephalopathy (mad cow disease) prion. Knowledge of this structure is vital in understanding the function of the protein. Structural information is usually classified as one of secondary, tertiary and quaternary structure.

One of the key ideas in bioinformatics is the notion of homology. In the genomic branch of bioinformatics, homology is used to predict the function of a gene: if the sequence of gene A, whose function is known, is homologous to the sequence of gene B, whose function is unknown, one could infer that B may share A's function. In the structural branch of bioinformatics, homology is used to determine which parts of a protein are important in structure formation and interaction with other proteins. In a technique called homology modeling, this information is used to predict the structure of a protein once the structure of a homologous protein is known. This currently remains the only way to predict protein structures reliably.

One example of this is hemoglobin in humans and the hemoglobin in legumes (leghemoglobin), which are distant relatives from the same protein superfamily. Both serve the same purpose of transporting oxygen in the organism. Although both of these proteins have completely different amino acid sequences, their protein structures are virtually identical, which reflects their near identical purposes and shared ancestor.

Other techniques for predicting protein structure include protein threading and de novo (from scratch) physics-based modeling.

Another aspect of structural bioinformatics include the use of protein structures for Virtual Screening models, such as Quantitative Structure-Activity Relationship models and ProteoChemometric Models (PCM). In addition, a protein's crystal structure can be used in simulation of for example ligand-binding studies and in silico mutagenesis studies.

11.2.1 Medical Informatics

The field of medical informatics or biomedical informatics includes the significant key applications in the following fields:

Molecular Medicine

Molecular medicine is a broad field, where physical, chemical, biological, bioinformatics and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them. The molecular medicine

perspective emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs. Basically the human genome has profound effects on the fields of biomedical research and clinical medicine. The completion of the human genome and the use of bioinformatic tools means that the researcher can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly. This new knowledge of the molecular mechanisms of disease will enable better treatments, cures and even preventative tests to be developed.

Personalised Medicine

Personalized medicine, also referred to as precision medicine, is a medical model that separates people into different groups—with medical decisions, practices, interventions and/or products being tailored to the individual patient based on their predicted response or risk of disease. The terms personalized medicine, precision medicine, stratified medicine and P4 medicine are used interchangeably to describe this concept though some authors and organisations use these expressions separately to indicate particular nuances.

While the tailoring of treatment to patients dates back at least to the time of Hippocrates, the term has risen in usage in recent years given the growth of new diagnostic and informatics approaches that provide understanding of the molecular basis of disease, particularly genomics. This provides a clear evidence base on which to stratify (group) related patients.

Among 14 Grand Challenges for Engineering, initiative sponsored by National Academy of Engineering (NAE), personalized medicine has been identified as a key and prospective approach to "Achieve Optimal Individual Health Decisions", therefore overcoming the challenge of "Engineer Better Medicines". Clinical medicine will become more personalised with the development of the field of pharmacogenomics. This is the study of how an individual's genetic inheritance affects the body's response to drugs.

In personalised medicine, diagnostic testing is often employed for selecting appropriate and optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. The use of genetic information has played a major role in certain aspects of personalized medicine (for example, pharmacogenomics), and the term was first coined in the context of genetics, though it has since broadened to encompass all sorts of personalization measures, including the use of proteomics, imaging analysis, nanoparticle-based theranostics, among others.

Every person has a unique variation of the human genome. Although most of the variation between individuals has no effect on health, an individual's health stems from genetic variation with behaviours and influences from the environment.

Modern advances in personalized medicine rely on technology that confirms a patient's fundamental biology, DNA, RNA, or protein, which ultimately leads to

confirming disease. For example, personalised techniques, such as genome sequencing can reveal mutations in DNA that influence diseases ranging from cystic fibrosis to cancer. Another method, called RNA-seq, can show which RNA molecules are involved with specific diseases. Unlike DNA, levels of RNA can change in response to the environment. Therefore, sequencing RNA can provide a broader understanding of a person's state of health. Recent studies have linked genetic differences between individuals to RNA expression, translation, and protein levels.

The concepts of personalised medicine can be applied to new and transformative approaches to health care. Personalised health care is based on the dynamics of systems biology and uses predictive tools to evaluate health risks and to design personalised health plans to help patients mitigate risks, prevent disease and to treat it with precision when it occurs. In some instances, personalised health care can be tailored to the mark-up of the disease causing agent instead of the patient's genetic mark-up; examples are drug resistant bacteria or viruses.

Preventative Medicine

Preventive medicine is a medical specialty recognized by the American Board of Medical Specialties (ABMS), which focuses on the health of individuals and communities. The goal of preventive medicine is to promote health and well-being and prevent disease, disability and death.

Preventive medicine specialists are licensed Medical Doctors (MD) or Doctors of Osteopathy (DO) who possess expertise in a broad range of health care skills, including biostatistics, epidemiology, planning and evaluation of health services, management of health care organizations, research, and the practice of prevention in clinical settings. They apply their knowledge and skills in medicine, social, economic, and behavioural sciences to improve the health and quality of life of individuals, families, communities and populations through disease prevention and health promotion.

Preventive medicine has three sub-specialty areas, namely the public health and general preventive medicine, the occupational medicine, and the aerospace medicine. With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a person's susceptibility to different diseases may become a distinct reality.

Preventive healthcare, or prophylaxis, consists of measures taken for disease prevention. Disease and disability are affected by environmental factors, genetic predisposition, disease agents, and lifestyle choices and are dynamic processes which begin before individuals realize they are affected. Disease prevention relies on anticipatory actions that can be categorized as primal, primary, secondary, and tertiary prevention.

There are many methods for prevention of disease. It is recommended that adults and children aim to visit their doctor for regular check-ups, even if they feel

NOTES

healthy, to perform disease screening, identify risk factors for disease, discuss tips for a healthy and balanced lifestyle, stay up to date with immunizations and boosters, and maintain a good relationship with a healthcare provider. Additionally, it is suggested that people must use the Air Quality Index (AQI) to check the level of pollution of inside air and outside air. Some common disease screenings include checking for hypertension (high blood pressure), hyperglycemia (high blood sugar, a risk factor for diabetes mellitus), hypercholesterolemia (high blood cholesterol), screening for colon cancer, depression, mammography (to screen for breast cancer), colorectal cancer screening, a Pap test (to check for cervical cancer), and screening for osteoporosis. Genetic testing can also be performed to screen for mutations that cause genetic disorders or predisposition to certain diseases such as breast or ovarian cancer.

Gene Therapy

Gene therapy is a medical field which focuses on the genetic modification of cells to produce a therapeutic effect or the treatment of disease by repairing or reconstructing defective genetic material.

Human gene therapy seeks to modify or manipulate the expression of a gene or to alter the biological properties of living cells for therapeutic use. Gene therapy is a technique that modifies a person's genes to treat or cure disease. Gene therapies can work by the below given mechanisms:

- Replacing a disease-causing gene with a healthy copy of the gene.
- Inactivating a disease-causing gene that is not functioning properly.
- Introducing a new or modified gene into the body to help treat a disease.

Gene therapy products are being studied to treat diseases including cancer, genetic diseases, and infectious diseases. Circular DNA molecules can be genetically engineered to carry therapeutic genes into human cells. Viruses have a natural ability to deliver genetic material into cells, and therefore some gene therapy products are derived from viruses. Once viruses have been modified to remove their ability to cause infectious disease, these modified viruses can be used as vectors (vehicles) to carry therapeutic genes into human cells. Bacteria can be modified to prevent them from causing infectious disease and then used as vectors (vehicles) to carry therapeutic genes into human tissues.

Human gene editing technology refers to gene editing and are specifically used to disrupt harmful genes or to repair mutated genes. The concept of gene therapy is to fix a genetic problem at its source. If, for instance, in an (usually recessively) inherited disease a mutation in a certain gene results in the production of a dysfunctional protein, gene therapy could be used to deliver a copy of this gene that does not contain the deleterious mutation, and thereby produces a functional protein. This strategy is referred to as gene replacement therapy and is employed to treat inherited retinal diseases.

NOTES

Drug Development

Drug development is the process of bringing a new pharmaceutical drug to the market once a lead compound has been identified through the process of drug discovery. It includes preclinical research on microorganisms and animals, filing for regulatory status, such as via the United States Food and Drug Administration (FDA) for an investigational new drug to initiate clinical trials on humans, and may include the step of obtaining regulatory approval with a new drug application to market the drug. The entire process – from concept through preclinical testing in the laboratory to clinical trial development, including Phase I–III trials – to approved vaccine or drug typically takes more than a decade. With an improved understanding of disease mechanisms and using computational tools to identify and validate new drug targets, more specific medicines that act on the cause, not merely the symptoms, of the disease can be developed. These highly specific drugs promise to have fewer side effects than many of today's medicines.

Microbial Genome Applications

The advent of the complete genome sequences and their potential to provide a greater insight into the microbial world and its capacities could have broad and far reaching implications for environment, health, energy and industrial applications. Consequently, in 1994, the US Department of Energy (DOE) initiated the MGP (Microbial Genome Project) to sequence genomes of bacteria useful in energy production, environmental cleanup, industrial processing and toxic waste reduction. By studying the genetic material of these organisms, scientists understand these microbes at a very fundamental level and isolate the genes that give them their unique abilities to survive under extreme conditions. For example, *Deinococcus radiodurans* is known as the world's toughest bacteria and it is the most radiation resistant organism known. Scientists are interested in this organism because of its potential usefulness in cleaning up waste sites that contain radiation and toxic chemicals. Additionally, the scientists are studying the genome of the microbe *Chlorobium tepidum* which has an unusual capacity for generating energy from light.

The archaeon *Archaeoglobus fulgidus* and the bacterium *Thermotoga maritima* have potential for practical applications in industry and government-funded environmental remediation.

Other industrially useful microbes include, *Corynebacterium glutamicum* which is of high industrial interest as a research object because it is used by the chemical industry for the biotechnological production of the amino acid lysine. The substance is employed as a source of protein in animal nutrition.

Biotechnologically produced lysine is added to feed concentrates as a source of protein, and is an alternative to soybeans or meat and bone meal. *Lactococcus lactis* is one of the most important microorganisms involved in the dairy industry.

NOTES

Scientists have been examining the genome of *Enterococcus faecalis* - a leading cause of bacterial infection among hospital patients. They have discovered a virulence region made up of a number of antibiotic-resistant genes that may contribute to the bacterium's transformation from a harmless gut bacteria to a menacing invader.

11.2.2 Cheminformatics and Pharmacoinformatics

Cheminformatics, also known as chemoinformatics, refers to use of physical chemistry theory with computer and information science techniques, sometimes called 'in silico' techniques, in application to a range of descriptive and prescriptive problems in the field of chemistry, including in its applications to biology and related molecular fields. Such in silico techniques are used, for example, by pharmaceutical companies and in academic settings to aid and inform the process of drug discovery, for instance in the design of well-defined combinatorial libraries of synthetic compounds, or to assist in structure-based drug design. The methods can also be used in chemical and allied industries, and such fields as environmental science and pharmacology, where chemical processes are involved or studied.

The term 'Chemoinformatics' was defined in its application to drug discover, for instance, by F.K. Brown in 1998, "Chemoinformatics is the mixing of those information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the area of drug lead identification and optimization".

Since then, both terms, cheminformatics and chemoinformatics, have been used, although, lexicographically, cheminformatics appears to be more frequently used despite academics in Europe declaring for the variant chemoinformatics in 2006.

Cheminformatics combines the scientific working fields of chemistry, computer science, and information science—for example in the areas of topology, chemical graph theory, information retrieval and data mining in the chemical space. Cheminformatics can also be applied to data analysis for various industries like paper and pulp, dyes and such allied industries.

A primary application of cheminformatics is the storage, indexing, and search of information relating to chemical compounds. The efficient search of such stored information includes topics that are dealt with in computer science, such as data mining, information retrieval, information extraction, and machine learning.

File Formats of Cheminformatics

The in silico representation of chemical structures uses specialized formats such as the Simplified Molecular Input Line Entry Specifications (SMILES) or the XML-based Chemical Markup Language. These representations are often used for storage in large chemical databases. While some formats are suited for visual representations in two- or three-dimensions, others are more suited for studying physical interactions, modeling and docking studies.

NOTES

Pharmacoinformatics

Drug discovery and development requires the integration of multiple scientific and technological disciplines. These include chemistry, biology, pharmacology, pharmaceutical technology and extensive use of information technology. The latter is increasingly recognised as Pharmacoinformatics. Pharmacoinformatics relates to the broader field of bioinformatics.

The main idea behind the field is to integrate different informatics branches, for example bioinformatics, chemoinformatics, immunoinformatics, etc., into a single platform, resulting in a seamless process of drug discovery. The first reference of the term 'Pharmacoinformatics' can be found in the year of 1993.

The first dedicated department for Pharmacoinformatics was established at the National Institute of Pharmaceutical Education and Research, S.A.S. Nagar, India in 2003. This has been followed by different universities worldwide including a program by European universities named the European Pharmacoinformatics Initiative (EuroPIN).

Definition of Pharmacoinformatics

Pharmacoinformatics is also referred to as pharmacy informatics. According to the article "***Pharmacy Informatics: What You Need to Know Now***" by the University of Illinois at Chicago Pharmacoinformatics may be defined as, "The scientific field that focuses on medication-related data and knowledge within the continuum of healthcare systems". It is the application of computers to the storage, retrieval and analysis of drug and prescription information. Pharmacy informaticists work with pharmacy information management systems that help the pharmacist safe decisions about patient drug therapies with respect to, medical insurance records, drug interactions, as well as prescription and patient information.

Pharmacy informatics can be thought of as a sub-domain of the larger professional discipline of health informatics. Health informatics is the study of interactions between people, their work processes and engineered systems within health care with a focus on pharmaceutical care and improved patient safety. For example, the Health Information Management Systems Society (HIMSS) defines pharmacy informatics as, "The scientific field that focuses on medication-related data and knowledge within the continuum of healthcare systems - including its acquisition, storage, analysis, use and dissemination - in the delivery of optimal medication-related patient care and health outcomes".

Check Your Progress

1. Explain the term bioinformatics.
2. What are the important sub-disciplines within bioinformatics and computational biology?
3. How the protein structure prediction is done?
4. What is molecular medicine?

11.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

NOTES

1. Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Common uses of bioinformatics include the identification of genes and Single Nucleotide Polymorphisms (SNPs).
2. Following are the important sub-disciplines within bioinformatics and computational biology:
    - Development and implementation of computer programs that enable efficient access to, management and use of, various types of information.
    - Development of new algorithms (mathematical formulas) and statistical measures that assess relationships among members of large data sets. For example, there are methods to locate a gene within a sequence, to predict protein structure and/or function, and to cluster protein sequences into families of related sequences.
3. Protein structure prediction is another important application of bioinformatics. The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the vast majority of cases, this primary structure uniquely determines a structure in its native environment. Evidently, there are exceptions, such as the bovine spongiform encephalopathy (mad cow disease) prion. Knowledge of this structure is vital in understanding the function of the protein. Structural information is usually classified as one of secondary, tertiary and quaternary structure.
4. Molecular medicine is a broad field, where physical, chemical, biological, bioinformatics and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them. The molecular medicine perspective emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs.

11.4 SUMMARY

- Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex.

NOTES

- As an interdisciplinary field of science, bioinformatics combines biology, computer science, information engineering, mathematics and statistics to analyse and interpret the biological data.
- Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Common uses of bioinformatics include the identification of genes and Single Nucleotide Polymorphisms (SNPs).
- In the field of genetics, bioinformatics aids in sequencing and annotating genomes and their observed mutations. It plays a role in the text mining of biological literature and the development of biological and gene ontologies to organize and query biological data. It also plays a role in the analysis of gene and protein expression and regulation.
- Bioinformatics tools aid in comparing, analysing and interpreting genetic and genomic data and more generally in the understanding of evolutionary aspects of molecular biology.
- In structural biology, it aids in the simulation and modeling of DNA, RNA, proteins as well as biomolecular interactions.
- The primary goal of bioinformatics is to enhance the understanding of biological processes. Examples include pattern recognition, data mining, machine learning algorithms, visualization, sequence alignment, gene finding, genome assembly, drug design, drug discovery, protein structure alignment, protein structure prediction, prediction of gene expression and protein–protein interactions, genome-wide association studies, the modeling of evolution and cell division/mitosis.
- In the context of genomics, annotation is the process of marking the genes and other biological features in a DNA sequence. This process needs to be automated because most genomes are too large to annotate by hand, not to mention the desire to annotate as many genomes as possible, as the rate of sequencing has ceased to pose a bottleneck.
- Annotation is made possible by the fact that genes have recognisable start and stop regions, although the exact sequence found in these regions can vary between genes.
- The expression of many genes can be determined by measuring mRNA levels with multiple techniques including microarrays, Expressed cDNA Sequence Tag (EST) sequencing, Serial Analysis of Gene Expression (SAGE) tag sequencing, Massively Parallel Signature Sequencing (MPSS), RNA-Seq, also known as 'Whole Transcriptome Shotgun Sequencing (WTSS)' or various applications of multiplexed in situ hybridization.
- The amino acid sequence of a protein, the so-called primary structure, can be easily determined from the sequence on the gene that codes for it. In the

NOTES

vast majority of cases, this primary structure uniquely determines a structure in its native environment. Evidently, there are exceptions, such as the bovine spongiform encephalopathy (mad cow disease) prion.

- Molecular medicine is a broad field, where physical, chemical, biological, bioinformatics and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them.
- The molecular medicine perspective emphasizes cellular and molecular phenomena and interventions rather than the previous conceptual and observational focus on patients and their organs.
- Basically the human genome has profound effects on the fields of biomedical research and clinical medicine. The completion of the human genome and the use of bioinformatic tools means that the researcher can search for the genes directly associated with different diseases and begin to understand the molecular basis of these diseases more clearly.
- Preventive medicine has three sub-specialty areas, namely the public health and general preventive medicine, the occupational medicine, and the aerospace medicine. With the specific details of the genetic mechanisms of diseases being unravelled, the development of diagnostic tests to measure a person's susceptibility to different diseases may become a distinct reality.
- Gene therapy is a medical field which focuses on the genetic modification of cells to produce a therapeutic effect or the treatment of disease by repairing or reconstructing defective genetic material.
- Human gene therapy seeks to modify or manipulate the expression of a gene or to alter the biological properties of living cells for therapeutic use. Gene therapy is a technique that modifies a person's genes to treat or cure disease.
- Cheminformatics, also known as chemoinformatics, refers to use of physical chemistry theory with computer and information science techniques, sometimes called 'in silico' techniques, in application to a range of descriptive and prescriptive problems in the field of chemistry, including in its applications to biology and related molecular fields.
- Pharmacoinformatics is also referred to as pharmacy informatics. According to the article "*Pharmacy Informatics: What You Need to Know Now*" by the University of Illinois at Chicago Pharmacoinformatics may be defined as, "The scientific field that focuses on medication-related data and knowledge within the continuum of healthcare systems". It is the application of computers to the storage, retrieval and analysis of drug and prescription information.

11.5 KEY WORDS

NOTES

- **Bioinformatics:** Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data, in particular when the data sets are large and complex. Bioinformatics has been used for in silico analyses of biological queries using mathematical and statistical techniques. Common uses of bioinformatics include the identification of genes and Single Nucleotide Polymorphisms (SNPs).
- **Molecular medicine:** Molecular medicine is a broad field, where physical, chemical, biological, bioinformatics and medical techniques are used to describe molecular structures and mechanisms, identify fundamental molecular and genetic errors of disease, and to develop molecular interventions to correct them.
- **Gene therapy:** Gene therapy is a medical field which focuses on the genetic modification of cells to produce a therapeutic effect or the treatment of disease by repairing or reconstructing defective genetic material. Gene therapy is a technique that modifies a person's genes to treat or cure disease.
- **Cheminformatics:** Cheminformatics, also known as chemoinformatics, refers to use of physical chemistry theory with computer and information science techniques, sometimes called 'in silico' techniques, in application to a range of descriptive and prescriptive problems in the field of chemistry, including in its applications to biology and related molecular fields.
- **Pharmacoinformatics:** Pharmacoinformatics is also referred to as pharmacy informatics. According to the article "*Pharmacy Informatics: What You Need to Know Now*" by the University of Illinois at Chicago Pharmacoinformatics may be defined as, "The scientific field that focuses on medication-related data and knowledge within the continuum of healthcare systems". It is the application of computers to the storage, retrieval and analysis of drug and prescription information.

11.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term bioinformatics.
2. Why medical informatics is important?
3. What are preventive medicines?
4. Explain the terms cheminformatics and pharmacoinformatics.

NOTES

Long-Answer Questions

1. Briefly discuss the concept of bioinformatics giving its features, fields where used with the help of appropriate examples.
2. Discuss in detail the significance of medical informatics in different medical fields giving relevant examples.
3. Explain the concept of gene therapy and drug development with reference to bioinformatics.
4. Elaborate briefly about the microbial genome applications giving relevant examples.
5. Discuss the basic features of cheminformatics and pharmacoinformatics.

11.7 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 12 CURRENT RESEARCHES IN BIOINFORMATICS

NOTES

Structure

12.0 INTRODUCTION

Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software. The key objective of bioinformatics is to enrich biological data and to apply computer based algorithm to analyse biological data. Bioinformatics helps in DNA barcoding; design the patterns of disease outbreaks and new biological products.

Bioinformatics today has come in every major discipline in biology. In genomics, bioinformatics has helped in genome sequencing, and has presented its success in locating the genes, in phylogenetic comparison and in the detection of transcription factor binding sites of the genes. Specific genes can be altered to improve the production of meat and milk. Positive changes can be made in their genome for disease resistant. Bioinformatics deals with the exponential growth in biological data have led to the development of primary and secondary databases of nucleic acid sequences, protein sequences and structures.

Bioinformatics applications in cancer have rapidly evolved over the past several years. Ever since its initial implementation, next generation sequencing has transformed the concept of cancer biology, and the approaches to analysing the more and more complex datasets have also become increasingly complex. Routine bioinformatics now range from those that rapidly detect and predict the functional impact of molecular alterations, to those that quantify the changes in the tumor microenvironment. Bioinformatics analysis can not only accelerate drug target identification and drug candidate screening and refinement, but also facilitate characterization of side effects and predict drug resistance.

In this unit, you will study about the current researches in bioinformatics and applications of bioinformatics in cancer detection and drug targets.

NOTES

12.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain about the current researches in bioinformatics
- Define the applications of bioinformatics in cancer detection and drug targets

12.2 BIOINFORMATICS: CURRENT RESEARCHES

Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.

Fundamentally, 'Bioinformatics' is an application of information technology for processing and analysing the data generated in biological researches and experiments or alternatively bioinformatics is a computerized system for molecular biology and has many practical's use in biology. The key objectives of bioinformatics is to enrich biological data and to apply computer based algorithm to analyse biological data. The analysis helps in DNA barcoding; designing the patterns of disease outbreaks and new biological products. Bioinformatics has high capacity to analysis data and to influence the attribute of plants, animals and human beings. The bioinformatics researches have processed to enhance the gene and protein sequences. In proteomics, bioinformatics helps in the study of protein structures and the discovery of sequence sites where protein-protein interactions happen.

At present, the bioinformatics has its implications in every major discipline in biology. In genomics, bioinformatics has helped in genome sequencing, and has presented its success in locating the genes, in phylogenetic comparison and in the detection of transcription factor binding sites of the genes. Bioinformatics provides analytical tools for microarray data. Analytical tools of bioinformatics provide relevant information about the genes that exist in the genome of these species. These tools help in predicting the function of different genes and factors affecting these genes. Positive changes can be made in their genome for disease resistant.

Paulien Hogeweg and Ben Hesper coined the term bioinformatics in 1970, and defined that bioinformatics is, "Study of Computer Processes in Biotic Systems". The information on information technology, computer science and biology is mainly handled by bioinformatics. Biologist conducts laboratory research and collects the gene expressions, DNA and protein sequences, etc. In the development of algorithms, tools, and data storage and analysis software, computer scientists are involved. By analysing molecular data with different programs and tools, bioinformaticians study biological problems.

Bioinformatics is used mainly in medicine, in microbial genome applications and in agriculture in a broad spectrum of applications.

NOTES

Modern uses of bioinformatics include genetic engineering and technology for cell and tissue culture. Biotechnology identified organisms and microorganisms which may be useful in the dairy sector and food processors in the field of bioinformatics. The *Lactococcus lactis*, a non-pathogenic rod shaped bacterium is unique and critical for the production of dairy products, such as buttermilk, yoghurt and cheese, is one of the most important microorganisms involved in the dairy sector. Bioinformatics researchers anticipate that the physiologic and genetic make-up of this bacterium to be invaluable for food producers as well as the L capability research pharmaceutical industry. Lactis can be used as a drug supply vehicle.

Latest Microbial Genome Applications in Bioinformatics

The significant applications of bioinformatics are used in the following fields of microbial genome applications.

Waste Cleanup

Bioinformatics identify the specific 'Bacteria' and 'Microbes' that help and support in the purification of wastes. *Deinococcus radioduran* bacterium has the capability to repair DNA damaged and small chromosome fragments by isolating the damaged segments within a concentrated area. *Deinococcus radioduran* is a bacteria which is listed in the World Book of Guinness. It has been used in organic chemicals, solvents, and heavy metals at sites of radioactive waste.

Climate Change

Climate change happens due to variations of the Earth's solar radiation, plate tectonics, and volcanic eruptions, or human changes to the natural world, including oceanic processes, such as the occurrence of oceanic circulation.

By studying microorganisms, genome researchers can understand these microbes at a very basic level, isolating genes that enable them to survive in extreme conditions. A phototrophic purple, a non-sulfur bacterium commonly found in soil and water is *Rhodopseudomonas palustris*. By absorbing carbon dioxide from the atmosphere and converting it into biomass, the sun is turned into cellular energy.

Biotechnology

Biotechnology typically comprises through breeding programming using the process of artificial selection and hybridization. The comprehensive concept of 'Biotechnology' covers a range of procedures for modifying live organisms in a manner that refers to animal domestication, plant cultivation and improvement.

Crop Improvement

Comparative genetics of plant genomes indicated that their genes' organization remained more conserved than had earlier been expected during evolutionary time. These findings specify that information from model crop systems can be used to suggest improvements to other food crops. Examples of the available complete plant genomes are *Arabidopsis thaliana* (watercress) and *Oryza sativa* (rice). The first sequenced plant *Arabidopsis thaliana* is considered as a model of plant genetics and biology investigation.

Improve Nutritional Quality

Recently, scientists have been able to transfer genes to rice to increase levels of Vitamin A, iron, and other micronutrients. This work might have a profound effect on the reduction of blindness and anaemia from vitamin A and iron deficiencies, respectively.

Scientists have inserted a yeast gene into the tomato and it turns out that a plant has a longer fruit life on the vine. Biologists have announced that the gene controlled by the humble fruit maturing process has been identified.

Medicine Applications

The following areas are being used in the field of medicine applications of bioinformatics.

Drug Discovery

When the first three-dimensional protein structure was determined the notion of using X-ray crystallography in drug discovery came up more than 30 years ago. In the space of a decade, a radical change has started in drug design, incorporating the know-how of 3D (Three Dimensional) target protein structures into the design process.

At every stage of the design process, protein structure can influence drug discovery. It is traditionally used in lead optimization, a process that uses a structure to guide the chemical modification of a lead molecule in order to optimize shape, hydrogen bonds, and other non-covalent interactions with the objective.

Personalized Medicine

Personalized medicine is a medical model that customizes health care with the application of genetic or other information to tailor all decisions and practices to the individual patient. Application outside long-established considerations, such as the family history of a patient, social conditions, the environment, and behaviour are very limited and almost no progress has been made.

Custom medical research search to identify solutions on the basis of each individual's susceptibility profile. The new diagnostic, medication development and individual therapy methods can be found.

Preventive Medicine

Prevention medicine or preventive treatment involves measures taken as precautionary measures and not to cure, to treat, or to treat diseases (or injuries). Simple examples include hand washing, nursing, and immunization for preventive medicine.

Gene Therapy

Gene therapy is a new form of drug delivery that includes a therapeutic agent produced by the patient's synthetic machinery. With the aim to produce enough protéin encoded by the gene (transgene), gene therapy involves the efficient introduction of functional genes into the patient's appropriate cells, so as to accurately and permanently correct the disorder.

12.2.1 Applications of Bioinformatics in Cancer Detection and Drug Targets

Bioinformatic applications in cancer have rapidly evolved over the past several years. Ever since its initial implementation, next generation sequencing has transformed the concept of cancer biology, and the approaches to analyse the more complex datasets. Routine bioinformatics pipelines currently range from those that rapidly detect and predict the functional impact of molecular alterations, to those that quantify the changes in the tumor microenvironment. For example, numerous tools that analyse tumor–immune interactions have been successfully developed for assessing the tumor infiltrating lymphocyte content, microsatellite instability, total mutational burden, and neoantigen presentation.

Some specific applications require high sensitivity, such as the quantification of tumor mutations from liquid biopsies (circulating cell free DNA). Additional novel applications try to enhance the capability to analyse the distribution and molecular impact of complicated genetic features, such as repetitive or transposable endogenous elements and exogenous genetic elements, for example human papilloma virus.

Cancer

Cancer is one of the leading causes of morbidity and mortality worldwide. There exists an urgent need to identify new biomarkers or signatures for early detection and prognosis. Mona *et al*. identified biomarker genes from functional network based on the 407 differential expressed genes between lung cancer and healthy populations from a public Gene Expression Omnibus dataset.

The lower expression of sixteen gene signature is associated with favorable lung cancer survival, DNA repair, and cell regulation. A new class of biomarkers, such as Alternative Splicing Variants (ASV) have been studied in recent years. Various platforms and methods, for example, Affymetrix Exon-Exon Junction Array, RNA-seq, and Liquid Chromatography Tandem Mass Spectrometry (LC–MS/MS), have been developed to explore the role of ASV in human disease.

Liquid Chromatography–Mass Spectrometry (LC–MS) is an analytical chemistry technique that combines the physical separation capabilities of Liquid Chromatography (or HPLC) with the mass analysis capabilities of Mass Spectrometry (MS). Coupled chromatography - MS systems are popular in chemical analysis because the individual capabilities of each technique are enhanced synergistically. This tandem technique can be used to analyse biochemical, organic, and inorganic compounds commonly found in complex samples of environmental and biological origin.

Cancer bioinformatics is a critical and important part of the systems clinical medicine in cancer and the core tool and approach to carry out the investigations of cancer in systems clinical medicine.

Cancer bioinformatics is expected to play a more important role in the identification and validation of biomarkers, specific to clinical phenotypes related

to early diagnoses, measurements to monitor the progress of the disease and the response to therapy, and predictors for the improvement of patient's life quality. Of gene-, protein-, peptide-, chemical- or physic-based variables in cancer, biomarkers were investigated from a single one to multiple markers, from the expression to functional indication, and from the network to dynamic network.

Network biomarkers as a new type of biomarkers with protein-protein interactions were investigated with the integration of knowledge on protein annotations, interaction, and signaling pathway. Alterations of network biomarkers can be monitored and evaluated at different stages and time points during the development of diseases, named dynamic network biomarkers, as one of the new strategies. Dynamic network biomarkers were expected to be correlated with clinical informatics, including patient complaints, history, therapies, clinical symptoms and signs, physician's examinations, biochemical analyses, imaging profiles, pathologies and other measurements.

Drug Targets

Most drug targets are members of families of proteins that are related phylogenetically. Examples include G-Protein Coupled Receptors (GPCRs), protein kinases, nuclear hormone receptors, serine proteases, and ion channels. The degree to which compounds that bind to the desired target also bind to these related proteins varies greatly and depends on the conservation of the protein fold and the sequence homology of the binding site.

A biological target is anything within a living organism to which some other entity (like an endogenous ligand or a drug) is directed and/or binds, resulting in a change in its behaviour or function. Examples of common classes of biological targets are proteins and nucleic acids. The definition is context-dependent, and can refer to the biological target of a pharmacologically active drug compound, the receptor target of a hormone (like insulin), or some other target of an external stimulus. Biological targets are most commonly proteins, such as enzymes, ion channels, and receptors.

The external stimulus, i.e., the drug or ligand physically binds to 'hits' the biological target. The interaction between the substance and the target may be:

- **Noncovalent** – A relatively weak interaction between the stimulus and the target where no chemical bond is formed between the two interacting partners and hence the interaction is completely reversible.
- **Reversible Covalent** – A chemical reaction occurs between the stimulus and target in which the stimulus becomes chemically bonded to the target, but the reverse reaction also readily occurs in which the bond can be broken.
- **Irreversible Covalent** – The stimulus is permanently bound to the target through irreversible chemical bond formation.

Depending on the nature of the stimulus, the following can occur:

- There is no direct change in the biological target, but the binding of the substance prevents other endogenous substances (such as, activating hormones) from binding to the target. Depending on the nature of the target, this effect is referred as receptor antagonism, enzyme inhibition, or ion channel blockade.
- A conformational change in the target is induced by the stimulus which results in a change in target function. This change in function can mimic the effect of the endogenous substance in which case the effect is referred to as receptor agonism (or channel or enzyme activation) or be the opposite of the endogenous substance which in the case of receptors is referred to as inverse agonism.

The term 'Biological Target' is frequently used in pharmaceutical research to describe the native protein in the body whose activity is modified by a drug resulting in a specific effect, which may be a desirable therapeutic effect or an unwanted adverse effect. In this context, the **biological target** is often referred to as a **drug target**. The most common drug targets of currently marketed drugs include:

- Proteins
  - o G Protein-Coupled Receptors (Target of 50% of Drugs)
  - o Enzymes (Especially Protein Kinases, Proteases, Esterases, and Phosphatases)
  - o Ion Channels
- ▪ Ligand-Gated Ion Channels
- ▪ Voltage-Gated Ion Channels
  - o Nuclear Hormone Receptors
  - o Structural Proteins, such as Tubulin
  - o Membrane Transport Proteins
- Nucleic Acids

Drug Target Identification

Identifying the biological origin of a disease, and the potential targets for intervention, is the first step in the discovery of a medicine using the reverse pharmacology approach. Potential drug targets are not necessarily disease causing but must by definition be disease modifying. An alternative means of identifying new drug targets is forward pharmacology based on phenotypic screening to identify 'Orphan' ligands whose targets are subsequently identified through target deconvolution.

Drug discovery starts with diagnosis of a disease with well characterized symptoms that reduce the quality of life. Conventionally, a desirable drug is a chemical (which could be a simple chemical or a complicated protein) or a combination of chemicals that reduces the symptoms without causing severe side

NOTES

effects in the patient. Other properties of a desirable drug include affordability and profit for drug companies, low chance of drug resistance leading to dramatic decrease in the commercial value of the drug, low deleterious effect on the environment, e.g., no re-activation by bacterial species after human use. Thus, a desirable drug is one that not only is efficacious with little side effects, but also has minimal long-term negative effect on the patient, the society and the environment.

Check Your Progress

1. Define the term bioinformatics.
2. What are the key objectives of bioinformatics?
3. Explain Liquid Chromatography–Mass Spectrometry (LC–MS).
4. What is biological target?

12.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.
2. The key objectives of bioinformatics is to enrich biological data and to apply computer based algorithm to analyse biological data. The analysis helps in DNA barcoding; designing the patterns of disease outbreaks and new biological products. Bioinformatics has high capacity to analysis data and to influence the attribute of plants, animals and human beings.
3. Liquid Chromatography–Mass Spectrometry (LC–MS) is an analytical chemistry technique that combines the physical separation capabilities of Liquid Chromatography (or HPLC) with the mass analysis capabilities of Mass Spectrometry (MS). Coupled chromatography - MS systems are popular in chemical analysis because the individual capabilities of each technique are enhanced synergistically. This tandem technique can be used to analyse biochemical, organic, and inorganic compounds commonly found in complex samples of environmental and biological origin.
4. A biological target is anything within a living organism to which some other entity (like an endogenous ligand or a drug) is directed and/or binds, resulting in a change in its behaviour or function. Examples of common classes of biological targets are proteins and nucleic acids. The definition is context-dependent, and can refer to the biological target of a pharmacologically active drug compound, the receptor target of a hormone (like insulin), or some other target of an external stimulus. Biological targets are most commonly proteins, such as enzymes, ion channels, and receptors.

12.4 SUMMARY

NOTES

- Bioinformatics is a field of study that uses computation to extract knowledge from biological data. It includes the collection, storage, retrieval, manipulation and modelling of data for analysis, visualization or prediction through the development of algorithms and software.
- The key objectives of bioinformatics is to enrich biological data and to apply computer based algorithm to analyse biological data. The analysis helps in DNA barcoding; designing the patterns of disease outbreaks and new biological products.
- Bioinformatics has high capacity to analysis data and to influence the attribute of plants, animals and human beings.
- The bioinformatics researches have processed to enhance the gene and protein sequences. In proteomics, bioinformatics helps in the study of protein structures and the discovery of sequence sites where protein-protein interactions happen.
- Analytical tools of bioinformatics provide relevant information about the genes that exist in the genome of these species. These tools help in predicting the function of different genes and factors affecting these genes. Positive changes can be made in their genome for disease resistant.
- Modern uses of bioinformatics include genetic engineering and technology for cell and tissue culture. Biotechnology identified organisms and microorganisms which may be useful in the dairy sector and food processors in the field of bioinformatics.
- The *Lactococcus lactis*, a non-pathogenic rod shaped bacterium is unique and critical for the production of dairy products, such as buttermilk, yoghurt and cheese, is one of the most important microorganisms involved in the dairy sector.
- Bioinformatics identify the specific 'Bacteria' and 'Microbes' that help and support in the purification of wastes. *Deinococcus radioduran* bacterium has the capability to repair DNA damaged and small chromosome fragments by isolating the damaged segments within a concentrated area.
- Bioinformatic applications in cancer have rapidly evolved over the past several years. Ever since its initial implementation, next generation sequencing has transformed the concept of cancer biology, and the approaches to analyse the more complex datasets.
- Some specific applications require high sensitivity, such as the quantification of tumor mutations from liquid biopsies (circulating cell free DNA).
- Additional novel applications try to enhance the capability to analyse the distribution and molecular impact of complicated genetic features, such as

NOTES

repetitive or transposable endogenous elements and exogenous genetic elements, for example human papilloma virus.

- Cancer is one of the leading causes of morbidity and mortality worldwide. There exists an urgent need to identify new biomarkers or signatures for early detection and prognosis.
- Mona *et al*. identified biomarker genes from functional network based on the 407 differential expressed genes between lung cancer and healthy populations from a public Gene Expression Omnibus dataset.
- The lower expression of sixteen gene signature is associated with favorable lung cancer survival, DNA repair, and cell regulation.
- A new class of biomarkers, such as Alternative Splicing Variants (ASV) have been studied in recent years. Various platforms and methods, for example, Affymetrix Exon-Exon Junction Array, RNA-seq, and Liquid Chromatography Tandem Mass Spectrometry (LC–MS/MS), have been developed to explore the role of ASV in human disease.
- Liquid Chromatography–Mass Spectrometry (LC–MS) is an analytical chemistry technique that combines the physical separation capabilities of Liquid Chromatography (or HPLC) with the mass analysis capabilities of Mass Spectrometry (MS).
- Coupled chromatography - MS systems are popular in chemical analysis because the individual capabilities of each technique are enhanced synergistically. This tandem technique can be used to analyse biochemical, organic, and inorganic compounds commonly found in complex samples of environmental and biological origin.
- Cancer bioinformatics is a critical and important part of the systems clinical medicine in cancer and the core tool and approach to carry out the investigations of cancer in systems clinical medicine.
- Cancer bioinformatics is expected to play a more important role in the identification and validation of biomarkers, specific to clinical phenotypes related to early diagnoses, measurements to monitor the progress of the disease and the response to therapy, and predictors for the improvement of patient's life quality.
- Of gene-, protein-, peptide-, chemical- or physic-based variables in cancer, biomarkers were investigated from a single one to multiple markers, from the expression to functional indication, and from the network to dynamic network.
- Network biomarkers as a new type of biomarkers with protein-protein interactions were investigated with the integration of knowledge on protein annotations, interaction, and signaling pathway.

- Dynamic network biomarkers were expected to be correlated with clinical informatics, including patient complaints, history, therapies, clinical symptoms and signs, physician's examinations, biochemical analyses, imaging profiles, pathologies and other measurements.
- Most drug targets are members of families of proteins that are related phylogenetically. Examples include G-Protein Coupled Receptors (GPCRs), protein kinases, nuclear hormone receptors, serine proteases, and ion channels.
- The degree to which compounds that bind to the desired target also bind to these related proteins varies greatly and depends on the conservation of the protein fold and the sequence homology of the binding site.
- A biological target is anything within a living organism to which some other entity (like an endogenous ligand or a drug) is directed and/or binds, resulting in a change in its behaviour or function. Examples of common classes of biological targets are proteins and nucleic acids.
- The definition is context-dependent, and can refer to the biological target of a pharmacologically active drug compound, the receptor target of a hormone (like insulin), or some other target of an external stimulus. Biological targets are most commonly proteins, such as enzymes, ion channels, and receptors.

12.5 KEY WORDS

- **Liquid Chromatography–Mass Spectrometry (LC–MS):** Liquid Chromatography–Mass Spectrometry (LC–MS) is an analytical chemistry technique that combines the physical separation capabilities of Liquid Chromatography (or HPLC) with the mass analysis capabilities of Mass Spectrometry (MS).
- **Network biomarkers:** Network biomarkers as a new type of biomarkers with protein-protein interactions were investigated with the integration of knowledge on protein annotations, interaction, and signaling pathway.
- **Dynamic network biomarkers:** Dynamic network biomarkers were expected to be correlated with clinical informatics, including patient complaints, history, therapies, clinical symptoms and signs, physician's examinations, biochemical analyses, imaging profiles, pathologies and other measurements.
- **Drug targets:** Most drug targets are members of families of proteins that are related phylogenetically. Examples include G-Protein Coupled Receptors (GPCRs), protein kinases, nuclear hormone receptors, serine proteases, and ion channels.

12.6 SELF-ASSESSMENT QUESTIONS AND EXERCISES

NOTES

Short-Answer Questions

1. What is bioinformatics?
2. Name some current researches in bioinformatics.
3. Which applications of bioinformatics is used in cancer detection?
4. What do you mean by drug targets?
5. Define the term biological target.

Long-Answer Questions

1. Briefly discuss the current researches in bioinformatics giving appropriate examples.
2. Discuss in detail the various applications of bioinformatics in cancer detection and drug targets.
3. Brief a note on latest microbial genome applications in bioinformatics.

12.7 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 13 ANIMAL GENOME DIVERSITY

NOTES

Structure

13.0 INTRODUCTION

Genome diversity is differ to the total number of genetic characteristics in the genetic makeup of a species, it ranges widely from the number of species to differences within species and can be attributed to the span of survival for a species. It is distinguished from genetic variability, which describes the tendency of genetic characteristics to vary. Genetic diversity serves as a way for populations to adapt to changing environments. With more variation, it is more likely that some individuals in a population will possess variations of alleles that are suited for the environment.

Deoxyribonucleic Acid (DNA) is a molecule composed of two polynucleotide chains that coil around each other to form a double helix carrying genetic instructions for the development, functioning, growth and reproduction of all known organisms and many viruses. DNA and Ribonucleic Acid (RNA) are Nucleic Acids. Alongside Proteins, Lipids and complex Carbohydrates (Polysaccharides), Nucleic Acids are one of the four major types of macromolecules that are essential for all known forms of life. The two DNA strands are known as Polynucleotides as they are composed of simpler monomeric units called Nucleotides. Each Nucleotide is composed of one of four Nitrogen-containing Nucleobases (Cytosine [C], Guanine [G], Adenine [A] or Thymine [T]), a sugar called Deoxyribose, and a Phosphate group.

NOTES

Biological databases also known as the libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including Genomics, Proteomics, Metabolomics, Microarray gene expression, and Phylogenetics. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures. Biological databases can be broadly classified into sequence, structure and functional databases. Nucleic acid and protein sequences are stored in sequence databases and structure databases store solved structures of RNA and proteins.

In this unit, you will study about the animal genome diversity, introduction to DNA and protein sequence analysis, introduction and concepts to biological database.

13.1 OBJECTIVES

After going through this unit, you will be able to:

- Explain the animal genome diversity
- Understand the DNA and protein sequence analysis
- Describe the biological database

13.2 ANIMAL GENOME DIVERSITY: BASICS FEATURES

Animals are highly diverse. Members of the animal kingdom are among the most conspicuous living things in the world. Animal evolution began in the ocean over 600 million years ago with tiny creatures that probably do not resemble any living organism today. Since then, animals have evolved into a highly diverse kingdom. Although over one million extant (currently living) species of animals have been identified, scientists are continually discovering more species as they explore ecosystems around the world. The number of extant species is estimated to be between 3 and 30 million.

But what is an animal? While we can easily identify dogs, birds, fish, spiders, and worms as animals, other organisms, such as corals and sponges, are not as easy to classify. Animals vary in complexity—from sea sponges to crickets to chimpanzees — and scientists are faced with the difficult task of classifying them within a unified system. They must identify traits that are common to all animals as well as traits that can be used to distinguish among related groups of animals. The animal classification system characterizes animals based on their anatomy, morphology, evolutionary history, features of embryological development, and genetic makeup. This classification scheme is constantly developing as new

information about species arises. Understanding and classifying the great variety of living species helps in understanding how to conserve the diversity of life on earth.

Animals are the eaters or consumers of the earth. They are heterotrophs and depend directly or indirectly on plants, photosynthetic Protists (Algae), or autotrophic bacteria for nourishment. Animals are able to move from place to place in search of food. In most, ingestion of food is followed by digestion in an internal cavity.

1. **Multicellular Heterotrophs:** All animals are multicellular heterotrophs. The unicellular heterotrophic organisms called *Protozoa,* which were at one time regarded as simple animals, are now considered to be members of the kingdom Protista, the large and diverse group.
2. **Diverse in Form:** Almost all animals (99%) are invertebrates, lacking a backbone. Of the estimated 10 million living animal species, only 42,500 have a backbone and are referred to as vertebrates. The animal kingdom includes about 35 phyla, most of which occur in the sea. Far fewer phyla occur in fresh water and fewer still occur on land. Members of the three phyla, namely the Arthropoda (spiders and insects), Mollusca (snails), and Chordata (vertebrates), dominate animal life on land.
3. **No Cell Walls:** Animal cells are distinct among multicellular organisms because they lack rigid cell walls and are usually quite flexible.
4. **Active Movement:** The ability of animals to move more rapidly and in more complex ways than members of other kingdoms is perhaps their most striking characteristic and one that is directly related to the flexibility of their cells and the evolution of nerve and muscle tissues.
5. **Sexual Reproduction:** Most animals reproduce sexually. Animal eggs, which are non-motile, are much larger than the small, usually flagellated sperm. In animals, cells formed in meiosis function directly as gametes. The haploid cells do not divide by mitosis first, as they do in plants and fungi, but rather fuse directly with each other to form the *Zygote*.
6. **Embryonic Development:** Most animals have a similar pattern of embryonic development. The zygote first undergoes a series of mitotic divisions, called *Cleavage,* and becomes a solid ball of cells, the morula, then a hollow ball of cells, the blastula. In most animals, the blastula folds inward at one point to form a hollow sac with an opening at one end called the blastopore. An embryo at this stage is called a *Gastrula*.

Genes are the small segments present on DNA. They encode for proteins which describes an individual's phenotype. They are characterized by the presence of start codon in the beginning and stop codon at the end. There are few genes which do not code for functional proteins and many are hypothetical due to their unknown functions. Interestingly, a gene may code for more than one proteins in

eukaryotes by the phenomenon known as *alternative splicing*. In short, genes are the basic unit of hereditary in all the organisms from simpler like prokaryotes to highly complex like eukaryotes. As we all must know by now that DNA consists of nucleic acids, sugars and phosphate, i.e., nucleotides which are of four types: Adenine, Guanine, Cytosine and Thymine. These nucleotide are arranged one after the other in a genome and encode for proteins which perform functions responsible for a phenotype. A codon is made up of three nucleotides which encodes for an amino acid. There are 64 possible codons and 20 amino acids, each amino acid is encoded by more than one codon. Also, three start and one stop codon is there which initiate or terminates the expression processes.

The term Gene was coined by Johannsen in 1909 and he named hereditary units of Mendel as 'Genes'. Subsequently many concepts and views emerged on genes calling them as the hereditary component, thread like structures ,etc. Basically, a gene encodes for a single polypeptide has a start codon in the beginning and stop codon at the end. Following points are under consideration of modern gene concept:

- Bothe male and female parents pass on their genes to the offspring, only Cytoplasmic inheritance is pass on from mother to offspring.
- DNA is arranged in chromosomes or linkage groups and in diploid organisms, genes are paired. The phenotype is the result of the behaviour resultant of their combination whether dominant or recessive, co-dominance, incomplete dominance, etc.
- Genes present on same chromosomes are generally transmitted together, called as linked genes.
- Gene has a position on chromosomes called locus. Due to chromosomal aberrations, locus may change which can affect phenotypic expression.
- Gamete formation leads to segregation of genes and haploid gametes have only one gene of its type.
- Non-linked genes assort differently and independently.

Now, as we understood the gene concept, let us focus on the regulation of gene expression in prokaryotes and eukaryotes.

The genomes of almost all living creatures, both plants and animals, consist of Deoxyribonucleic Acid (DNA), the chemical chain that includes the genes that code for different proteins and the regulatory sequences that turn those genes on and off.

Gene delivery refers to the process of introducing foreign genetic material, such as Deoxyribonucleic Acid (DNA) or Ribonucleic Acid (RNA) directly into host cells. Genetic material either DNA or RNA must reach the genome of the host cell in order to induce gene expression. However, this is not the end of process; successful gene delivery must ensure the foreign genetic material to stay stable

within the host cell. Also, the foreign material either integrates into the genome or replicate independently of it.

NOTES

Genomic diversity is a driving force influencing human and animal health, and susceptibility to disease. During the Keystone Symposium on Leveraging Genomic Diversity to Promote Human and Animal Health held in Kampala on Lake Victoria in Uganda, we brought together diverse communities of geneticists with primary objectives to explore areas of common interest, joint technological and methodological developments and applications, and to leverage opportunities for cross-learning. We explored translational genomics research in farmed animals and humans, debated the differences in research objectives in high- and low-resourced environments, delved into infectious diseases and zoo noses affecting humans and animals and considered diversity and cultural context at many levels. The 109 participants were from 22 countries (13 in Africa) and included 44 global travel awardees from 9 countries, equal numbers of men and women, of whom 31 were students and 13 senior investigators.

Animal genetics and breeding are focused on the performance of single or groups of animals against predetermined and accurately measurable parameters including but not limited to productivity, adaptation/resilience, and growth, tolerance/susceptibility to diseases, feed conversion and energy utilisation. This is achieved through the selection of desirable traits, a process that has developed over several millennia, since the domestication of animals, and the field has experienced dramatic acceleration over the past two decades with the advent of the genomics era, as explained by Michel Georges. There have been remarkable successes in animal breeding, leading to considerable genetic improvement and gains supporting productivity (e.g. milk, meat and egg production), adaptation, growth and other key traits. These significant genetic gains have also brought about an added environmental bonus in the form of a reduced carbon footprint.

Since our understanding of genome–phenome correlation is premised on good data, there were debates about the most suitable technologies to provide us with accurate reference genomes to stimulate the discovery of novel insights and a more accurate understanding of the biological mechanisms. Long-range sequencing and single-molecule sequencing technologies are coming of age and promise to provide more accurate and complete reference genomes. This will ensure that the genomes of all species are more fully characterized, especially with regard to Copy-Number Variants (CNVs) and other structural variations that are predicted to have a profound impact on biological processes.

Evolution and genome adaptation were a constant refrain throughout the sessions. Evan Eichler presented the Keynote Address on the relevance of primate evolution to human and animal health. He explained that segmental duplications, ranging upwards from 50bp (with CNVs now having been redefined to a lower threshold from the previous 1000-bp limit) have been understudied in most species as they are often intractable when analysing short-read sequences. The duplication

events can occur within chromosomes (intra-segmental duplications) or between chromosomes (inter-segmental duplications) and create regions of potential instability and hotspots for non-homologous crossover events.

NOTES

In the world of animal genetics and breeding, there are good examples of monogenic traits or large-effect variants that have been sought by breeders (e.g. polled cattle, hair coverage and length and Porcine Reproductive and Respiratory Syndrome (PRRS) resistance in pigs, as explained by Simon Lillico). Attempts to establish the desired alleles to fixation in herds have sometimes been hampered by the hitch-hiking of undesirable traits. Genome-editing technologies are now being considered as they target only the gene of interest, but this approach has led to safety and ethical concerns among sectors of the public as illustrated in the presentation by Alison van Eenennaam. It is, however, a potentially revolutionary technology for boosting agricultural production and lowering costs and adverse environmental impacts.

In the animal world, it is common to refer to breeding value and preservation of germplasm of individuals and their potential impact on the development of the herd. John Hickey led us through the process of how a deep understanding of the relationship between the genetic variants and the phenotype can be applied to the notion that an embryo's potential value can be determined even before the phenotype is fully formed. Although there are important monogenic traits relevant to animal breeding, most desirable traits are polygenic and multifactorial, like milk production and fat content in cattle. The shorter generation time and control over the environment, for example in terms of diet, temperature and exposures, make it more tractable to assess and predict the phenotypic outcome among animals. David Evans provided excellent explanations on how to assess genetic causality in the context of complex traits. Susan Lamont spoke about genetic associations for traits in poultry breeding and Mark Fife on the diversity of immune loci in Europe and Africa.

Genetic diversity is important because it helps maintain the health of a population, by including alleles that may be valuable in resisting diseases, pests and other stresses. If the environment changes, a population that has a higher variability of alleles will be better able to evolve to adapt to the new environment.

13.3 INTRODUCTION TO DEOXYRIBONUCLEIC ACID (DNA) AND PROTEIN SEQUENCE ANALYSIS

Deoxyribose Nucleic Acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying the genetic instructions used in the growth, development, functioning, are reproduction of all known living organisms and many viruses. DNA and Ribose Nucleic Acid (RNA) are Nucleic Acids; alongside Proteins, Lipids and complex Carbohydrates (Polysaccharides),

Nucleic Acids are one of the four major types of macromolecules that are essential for all known forms of life.

The two DNA strands are also known as polynucleotides as they are composed of simpler monomeric units called Nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases (Cytosine[C], Guanine [G], Adenine [A] or Thymine [T]), a sugar called Deoxyribose, and a Phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. The nitrogenous bases of the two separate polynucleotide strands are bound together, according to base pairing rules (A with T and C with G), with Hydrogen Bonds to make double-stranded DNA.

History

Nucleic acids were first isolated by Friedrich Miescher (1869) from pus cells. They were named nuclein. Hertwig (1884) proposed nuclein to be the carrier of hereditary traits. Because of their acidic nature they were named *nucleinic acids* and then nucleic acids (Altmann, 1899).

Fisher (1880s) discovered the presence of purine and pyri-midine bases in nucleic acids. Levene (1910) found deoxyribose nucleic acid to contain phosphoric acid as well as deoxyribose sugar.

He characterised four types of *Nucleotides* present in DNA. In 1950, Chargaff found that purine and pyrimidine content of DNA was equal. By this time W.T. Astbury had found through X-ray diffraction that DNA is a polynucleotide with nucleotides arranged perpendicular to the long axis of the molecule and separated from one another by a distance of 0.34 nm.

In 1953, Wilkins and Franklin got very fine X-ray photographs of DNA. The photographs showed that DNA was a helix with a width of 2.0 nm. One turn of the helix was 3.4 nm with 10 layers of bases stacked in it. Watson and Crick (1953) worked out the first correct double helix model from the X-ray photo-graphs of Wilkins and Franklin. Wilkins, Watson and Crick were awarded Nobel Prize for the same in 1962.

Watson and Crick (1953) built a 3D, molecular model of DNA that satisfied all the details obtained from X-ray photographs. They proposed that DNA consisted of a double helix with two chains having sugar phosphate on the outside and nitrogen bases on the inner side.

The *nitrogen bases* of the two chains formed complementary pairs with purine of one and pyrimidine of the other held together by Hydrogen Bonds (A-T, C-G). Comple-mentary base pairing between the two polynucleotide chains is considered to be hall mark of their proposition. It is of course based on early finding of Chargaff that A = T and Ñ = G Their second big proposal was that the two chains are antiparallel with $5' \rightarrow 3'$ orientation of one and $Y \rightarrow 5'$ orientation of the other.

NOTES

The two chains are twisted helically just as a rope ladder with rigid steps twisted into a spiral. Each turn of the spiral contains 10 nucleotides. This double helix or duplex model of DNA with antiparallel polynucleotide chains having complementary bases has an implicit mechanism of its replication and copying.

Here both the polynucleotide chains function as templates forming two double helices, each with one parent chain and one new but complementary strand. The phenomenon is called *semi conservative replication*. In-vitro synthesis of DNA has been carried out by Kornberg in 1959.

Types of DNA

DNA duplex model proposed by **Watson and Crick** is right handed spiral and is called B-DNA (Balanced DNA). In the model the base pairs lie at nearly right angles to the axis of helix. Another right handed duplex model is A-DNA (Alternate DNA). Here, a single turn of helix has 11 base pairs.

The base pairs lie 20° away from perpendicular to the axis. C-DNA has 9 base pairs per turn of spiral while in D-DNA the number is only 8 base pairs. Both are right handed. Z-DNA (Zigzag DNA) is left-handed double helix with zigzag back-bone, alternate *purine and pyrimidine bases*, single turn of 45 Å length with 12 base pairs and a single groove (Refer Figure 13.1).

B-DNA is more hydrated and most frequently found DNA in living cells. It is physiologically and biologically active form. However, it can get changed into other forms. Right handed DNA is known to change temporarily into the left handed form at least for a short distance. Such changes may cause changes in gene expression.

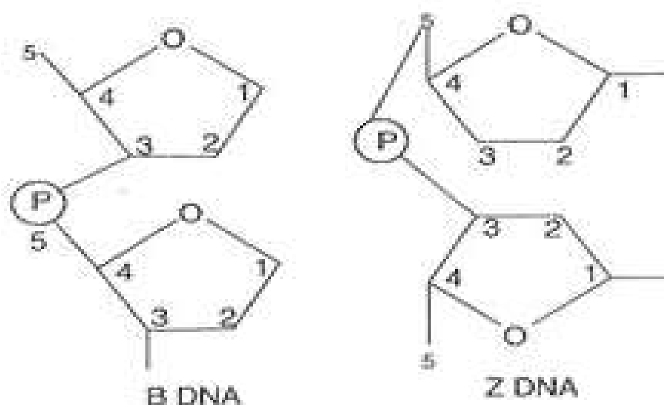


Fig. 13.1 *Orientation of Adjacent Sugar Molecules in B and Z DNA*

Circular and Linear DNA

In many prokaryotes the two ends of a DNA duplex are covalently linked to form circular DNA. Circular DNA is naked, that is, without association with histone proteins, though polyamines do occur. In linear DNA the two ends are free. It is found in eukaryotic nuclei where it is associated with histone proteins.

Linear DNA, without association with histone proteins, also occurs in some prokaryotes, for example Myco-plasma. In semi-autonomous cell organelles (mitochondria, plastids) DNA is circular, less commonly linear. It is always naked.

Chargaff's Rules

Chargaff (1950) made observations on the bases and other components of DNA. These observations or generalisations are called ***Chargaff's base equivalence rule***.

(i) Purine and pyrimidine base pairs are in equal amount, that is, Adenine + Guanine = Thymine + Cytosine. [A + G] = [T + C], i.e., [A+G] / [T+C] = 1

(ii) Molar amount of adenine is always equal to the molar amount of Thymine. Similarly, molar concentration of Guanine is equalled by molar concentration of Cytosine.

[A] = [T], i.e., [A] / [T] = 1; [G] = [C], i.e., [G] / [C] = 1

(iii) *Sugar deoxyribose* and *phosphate* occur in equimolar proportions.

(iv) A-T base pairs are rarely equal to Ñ—G base pairs.

(v) The ratio of [A+T] / [G+C] is variable but constant for a species (Refer Table 13.1). It can be used to identify the source of DNA. The ratio is low in primitive organisms and higher in advanced ones.

Table 13.1 *The Ratio of [A+T] / [G+C]*

Species	**A**	**G**	**C**	**T**	**A+T/ C+G**
1. Man	30.4	19.0	19.9	30.1	1.55
2. Calf	29.0	21.2	21.2	28.5	1.35
3. Wheat germ	28.1	21.8	22.7	27.4	1.25
4. Pea	30.8	19.2	18.5	30.5	1.62
5. Euglena	22.6	27.7	25.8	24.4	0.88
6. Escherichia coli	24.7	26.0	25.7	23.6	0.93

NOTES

13.3.1 Structure of DNA

DNA or Deoxyribose Nucleic Acid is a helically twisted double chain *poly deoxy ribonucleotide macromolecule* which constitutes the genetic material of all organisms with the exception of rhinoviruses. In prokaryotes it occurs in nucleoid and plasmids. This DNA is usually circular. In eukaryotes, most of the DNA is found in chromatin of nucleus.

It is linear, in smaller quantities of circular, double stranded DNA are found in *mitochondria* and *plastids* (organelle DNA). Small sized DNAs occur in viruses $\phi\times174$ bacteriophage has 5386 nucleotides. Bacteriophage lambda (Phage X) possesses 48502 Base Pairs (BP) while number of base pairs in Escherichia coli is 4.6 x 10^6. A single genome (haploid set of 23 chromosomes) has about 3.3 x 10^9 bp in human beings. Single-stranded DNA occurs as a genetic material in some viruses, for example phage $\phi \times 174$, coli phage fd, M13. DNA is the largest macromolecule with a diameter of 2 nm (20Å or 2×10^{-9}m) and often having 3 length in millimetres.

It is negatively charged due to phosphate groups. It is a long chain polymer of generally several hundred thousands of deoxy ribonucleotides. A DNA molecule has two un-branched comple-mentary strands. They are spirally coiled. The two spiral strands of DNA are collectively called DNA duplex (Refer Figure 13.2).

The two strands are not coiled upon each other but the whole double strand (DNA duplex) is coiled upon itself around a common axis like a rope stair case with solid steps twisted into a spiral. Due to spiral twisting, the DNA duplex comes to have two types of alternate grooves, major (22 Å) and minor (12 Å).

In B-DNA, one turn of the spiral has about 10 nucleotides on each strand of DNA. It occupies a distance of about 3.4 nm (34 Å or 3.4×10^{-9} m) so that adjacent nucleotides or their bases are separated by a space of about 0.34 nm (0.34×10^{-9} m or 3.4 Å).

A Deoxy Ribonucleotide of DNA is formed by cross-linking of three chemicals Ortho- Phosphoric Acid (H_3PO_4), Deoxyribose Sugar ($C_5H_{10}O_4$) and a nitrogen base. Four types of nitrogen bases occur in DNA. They belong to two groups, purines (9-membered double rings with nitrogen at 1, 3, 7 and 9 positions) and pyrimidines (six membered rings with nitrogen at 1 and 3 positions). DNA has two types of purines (Adenine or A and Guanine or G) and two types of pyrimidines (Cytosine or Ñ and Thymine or T).

NOTES



Fig. 13.2 *Structure of DNA*

Depending upon the type of nitrogen base, DNA has four kinds of deoxy ribonucleotides —deoxy Adenosine 5- Monophosphate (d AMP), deoxy Guaninosine 5-Monophosphate (d GMP), deoxy Thymidine 5-Monophosphate (d TMP) and deoxy Cytidine 5- Monophosphate (d CMP).

The back bone of a DNA chain or strand is built up of alternate deoxyribose sugar and phosphoric acid groups. The phosphate group is connected to carbon 52 of the sugar residue of its own nucleotide and carbon Ó of the sugar residue of the next nucleotide by (3'→ 5') phosphodiester bonds. -H of phosphate and -OH of sugar are eliminated as H_{20} during each ester formation.

NOTES

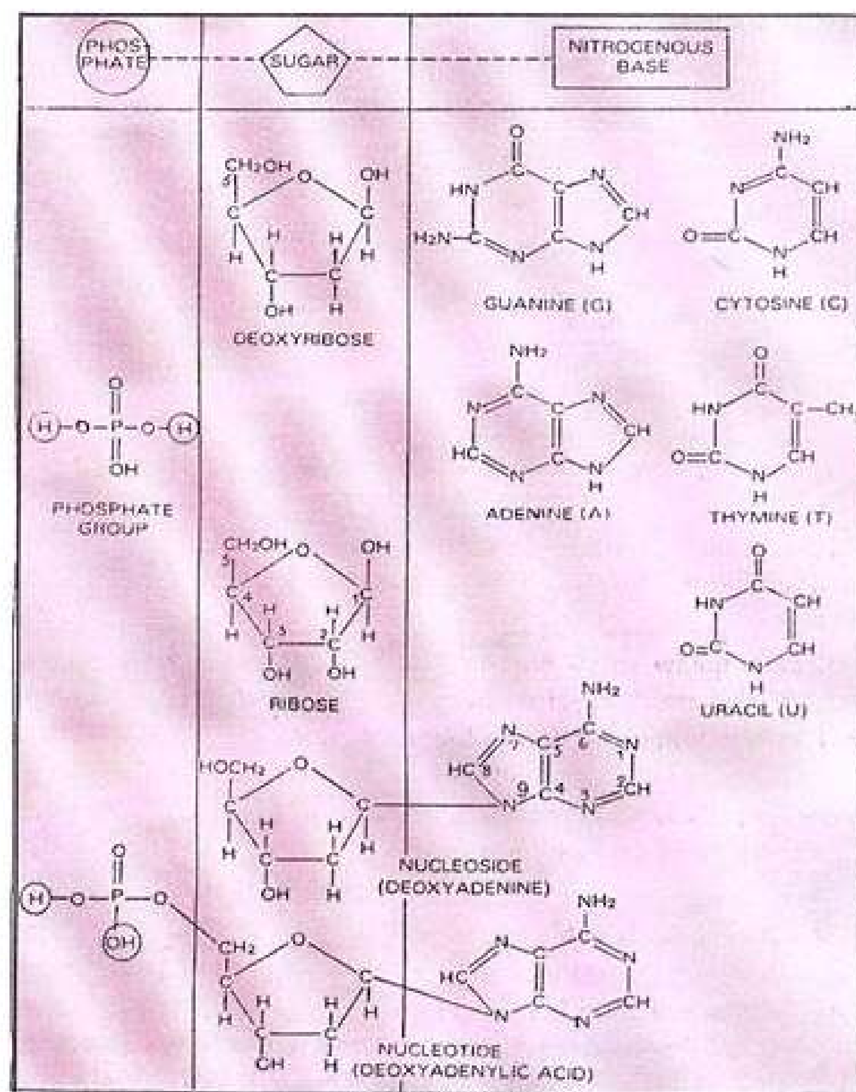


Fig. 13.3 *Building Blocks of Nucleotides*

Phosphate group provides acidity to the nucleic acids because at least one of its side group is free to dissociate. *Nitrogen bases* lie at right angles to the longitudinal axis of DNA chains. They are attached to carbon atom 1 of the sugars by N-Glycosidic bonds. Pyrimidine (C or T) is attached to deoxyribose by its

N-atom at 1 position while a purine (A or G) does so by N-atom at 9 position.

The two DNA chains are antiparallel that is, they run parallel but in opposite directions. In one chain the direction is $5' \rightarrow Y$ while in the opposite one it is $3' \rightarrow 5'$ (Refer Figure 13.3). The two chains are held together by hydrogen bonds between their bases. Adenine (A), a purine of one chain lies exactly opposite Thymine (T), a Pyrimidine of the other chain. Similarly, Cytosine (C, a Pyrimidine) lies opposite Guanine (G a purine). This allows a sort of lock and key arrangement between large sized purine and small sized pyrimidine.

It is strengthened by the appearance of hydrogen bonds between the two. Three hydrogen bonds occur between Cytosine and Guanine (C = G) at positions 1′–1′, 2′ – 6′ and 6′ – 2′. There are two such hydrogen bonds between adenine and thymine (A=T) which are formed at positions 1′– 3′ and 6′ – 4′. Hydrogen bonds occur between hydrogen of one base and oxygen and nitrogen of the other base. Since specific and different nitrogen bases occur on the two DNA chains, the latter are complementary.

Thus the sequence of say AAGCTCAG of one chain would have a complementary sequence of TTCGAGTC on the other chain. In other words, the two DNA chains are not identical but complementary to each other. It is because of specific base pairing with a purine lying opposite a pyrimidine. This makes the two chains 2 nm thick.

A purine- purine base pair will make it thicker while a pyrimidine- pyrimidine base pair will make it narrower than 2 nm. Further, A and Ñ or G and T do not pair because they fail to form hydrogen bonds between them. 5′ end of each chain bears phosphate radical while the 3′ end possesses a sugar residue (3′ – OH).

Salient Features of B model of DNA of Watson and Crick

- DNA is the largest biomolecule in the cell.
- DNA is negatively charged and dextrorotatory.
- Molecular configuration of DNA is 3D.
- DNA has two polynucleotide chains.
- The two chains of DNA have antiparallel polarity, 5′ → 3′ in one and 3′→ 5′ in other.
- Backbone of each polynucleotide chain is made of alternate *sugar-phosphate* groups. The nitrogen bases project inwardly.
- Nitrogen bases of two polynucleotide chains form complementary pairs, An opposite T and C opposite G.
- A large sized purine always comes opposite a small sized pyrimidine. This generates uniform distance between two strands of helix.
- Chain by two hydrogen bonds. Cytosine (C) of one chain is similarly held to Guanine of the other chain by three hydrogen bonds.
- The double chain is coiled in a helical fashion. The coiling is right handed. This coiling produces minor and major grooves alternately.
- The pitch of helix is 3.4 nm (34 A) with roughly 10 base pairs in each turn. The average distance between two adjacent base pairs comes to about 0.34 nm (0.34×10^{-9} m or 3.4 A).
- Planes of adjacent base pairs are stacked over one another. Along with *hydrogen bonding*, the stacking confers stability to the helical structure.

- DNA is acidic. For its compaction, it requires basic (histone) proteins. The histone proteins are +ve charged and occupy the major grooves of DNA at an angle of 30° to helix axis.

NOTES

Sense and Antisense Strands

Both the strands of DNA do not take part in controlling heredity and metabolism. Only one of them does so. The DNA strand which functions as template for RNA synthesis is known as template strand, minus (-) strand or antisense strand.

Its complementary strand is named nontemplate strand, plus (+) strand, sense and coding strand. The latter name is given because by convention DNA genetic code is written according to its sequence.

(5′) GCATTCGGCTAGTAAC (3′)

DNA Nontemplate, Sense (+) or Coding Strand

(3′) CGTAAGCCGATCATTG (5′)

DNA Template, Antisense, or Noncoding or (–) Strand

(5′) GCAUUCGGCUAGUAAC (3′)

RNA Transcript

RNA is transcribed on $3' \rightarrow 5'$ (-) strand (template/anti strand) of DNA in $5 \rightarrow 3$ direction. The term antisense is also used in wider prospective for any sequence or strand of DNA (or RNA) which is complementary to mRNA.

Denaturation and Renaturation

The H-bonds between nitrogen bases of two strands of DNA can break due to high temperature (82-90°C) or low or high pH, so that the two strands separate from each other. It is called *denaturation or melting*. Since A-T base pair has only 2H bonds, the area rich in A-T base pairs can undergo easy denaturation (melting). These areas are called low melting areas because they denature at comparatively low temperature. The area rich in G- C base pairs (called high melting area) is comparatively more stable and dense because three hydrogen bonds connect the G-C bases.

These areas have high Temperature of melting (Tm). On melting the viscosity of DNA decreases. The denatured DNA has the tendency to re-associate, i.e., the DNA strands separated by melting at 82-90°C can re-associate and form duplex on cooling to temperature at 65°C. It is called *renaturation or annealing*.

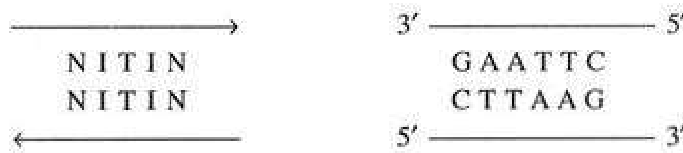
Denatured or separated DNA strands absorb more light energy than the intact DNA double strand. The increased absorption of light energy by separated or denatured DNA strands is called hyperchromatic effect. The effect is used in knowing whether DNA is single or double stranded.

Palindromic DNA

DNA duplex possesses areas where sequence of nucleotides is the same but opposite in the two strands. These sequences are recognised by restriction

endonucleases and are used in genetic engineering. Given hereunder sequence of bases in one strand ($3' \rightarrow 5'$) is GAATTC. It is same in other strand when read in $5' \rightarrow 3'$ direction. It is similar to palindrome words having same words in both forward and backward direction, for example NITIN, MALAYALAM.

NOTES



Repetitive DNA

It is the DNA having multiple copies of identical sequences of nitrogen bases. The number of copies of the same base sequence varies from a few to millions. DNA having single copy of base sequences is called unique DNA. It is made of functional genes. rRNA genes are, however, repeated several times. Repetitive DNA may occur in tandem or inter-spersed with unique sequences.

It is of two types, highly repetitive and moderately repetitive. Highly repetitive DNA consists of short sequences of less than 10 base pairs which are repeated millions of times. They occur in preeentromeric regions, heterochromatic re-gions of Y-chromosomes and satellite regions. Moderately repetitive DNA consists of a few hundred base pairs repeated at least 1000 times. It occurs in telomeres, centromeres and transposons.

Tandemly repeated sequences are especially liable to undergo misalignments during chromosome pairing, and thus the size of tandem clusters tends to be highly poly-morphic, with wide variations between individuals. Smaller clusters of such sequences can be used to characterize individual genomes in the technique of 'DNA-finger-printing'.

Satellite DNA

It is that part of repetitive DNA which has long repetitive nucleotide sequences in tandem that forms a separate fraction on density ultra-centrifugation upon the number of base pairs involved in repeat regions, satellite DNA is of two types, microsatellite sequences (1-6 bp repeat units flanked by conserved sequences) and minisatellite sequences (11-60 bp flanked by conserved restriction sites). The latter are hyper variable and are specific for each individual. They are being used for DNA matching or finger printing as first found out by Jeffreys et al (1985).

Genetic Information

The arrangement of nitrogen bases of DNA (and its product mRNA) determines the sequence of amino acid groups in polypeptides or proteins formed over ribosomes. One amino acid is specified by the sequence of three adjacent nitrogen bases. The latter is called *codon*. The segment of DNA which determines the synthesis of complete polypeptide is known as *Cistron*.

In procaryotes, a cistron has a continuous coding sequence from beginning to end. In eucaryotes a cistron contains noncoding regions which do not produce

part of gene product. They are called introns. Introns are often variable. The coding parts are known as exons. Cistrons having introns are *called split genes*.

Coding and Noncoding DNA

NOTES

Depending on the ability to form functional or non-functional products, DNA has two types of segments, coding and noncoding. In eukaryotes a greater part of DNA is noncoding since it does not form any functional product. They often possess repeated sequences or repetitive DNA. Most of them have fixed positions.

Some can move from one place to another. The mobile sequences are called jumping genes or transposons. In prokaryotes the amount of noncoding or non-functional DNA is small. Coding DNA consists of coding DNA sequences. These are of 2 types — protein coding sequences coding for all proteins except histone and non-protein coding sequences for tRNA, rRNA and histones.

13.3.2 Following the Functions of DNA

- **Genetic Information (Genetic Blue Print):** DNA is the genetic material which car-ries all the hereditary information. The genetic information is coded in the arrangement of its nitrogen bases.
- **Replication:** DNA has unique property of replication or production of carbon copies (Autocatalytic function). This is essential for transfer of genetic information from one cell to its daughters and from one generation to the next.
- **Chromosomes:** DNA occurs inside chromosomes. This is essential for equitable distribution of DNA during cell division.
- **Recombinations:** During meiosis, crossing over gives rise to new combination of genes called recombinations.
- **Mutations:** Changes in sequence of nitrogen bases due to addition, deletion or wrong replication give rise to mutations. Mutations are the fountain head of all variations and evolution.
- **Transcription:** DNA gives rise to RNAs through the process of transcription. It is heterocatalytic activity of DNA.
- **Cellular Metabolism:** It controls the metabolic reactions of the cells through the help of specific RNAs, synthesis of specific proteins, enzymes and hormones.
- **Differentiation:** Due to differential functioning of some specific regions of DNA or genes, different parts of the organisms get differentiated in shape, size and functions.
- **Development:** DNA controls development of an organism through working of an internal genetic clock with or without the help of extrinsic information.
- **DNA Finger Printing:** Hypervariable microsatellite DNA sequences of each indi-vidual are distinct. They are used in identification of individuals

and deciphering their relationships. The mechanism is called DNA finger printing.

- **Gene Therapy:** Defective heredity can be rectified by incorporating correct genes in place of defective ones.
- **Antisense Therapy:** Excess availability of anti-mRNA or antisense RNAs will not allow the pathogenic genes to express themselves. By this technique failure of angioplasty has been checked. A modification of this technique is RNA interference (RNAi).

NOTES

13.3.3 Protein Sequence in DNA

DNA sequencing enables us to perform a thorough analysis of DNA because it provides us with the most basic information of all: the sequence of nucleotides. With this knowledge, for example we can locate regulatory and gene sequences, make comparisons between homologous genes across species and identify mutations. Scientists recognised that this could potentially be a very powerful tool, and so there was competition to create a method that would sequence DNA. Then in 1974, two methods were independently developed by an American team and an English team to do exactly this. The Americans, led by Maxam and Gilbert, used a '*Chemical Cleavage Protocol'*, while the English, led by Sanger, and designed a procedure similar to the natural process of DNA replication. Even though both teams shared the 1980 Nobel Prize, Frederick Sanger's method became the standard because of its practicality (Speed, 1992). Sanger's method, which is also referred to as dideoxy sequencing or chain termination, is based on the use of dideoxy Nucleotides (ddNTP's) in addition to the normal nucleotides (NTP's) found in DNA. Dideoxynucleotides are essentially the same as nucleotides except them contain a hydrogen group on the 3' carbon instead of a Hydroxyl Group (OH). These modified nucleotides, when integrated into a sequence, prevent the addition of further nucleotides. This occurs because a phosphodiester bond cannot form between the dideoxynucleotide and the next incoming nucleotide, and thus the DNA chain is terminated (Refer Figure 13.4).

Method of Sanger Sequencing

The DNA sample to be sequenced is combined in a tube with primer, DNA polymerase, and DNA nucleotides (dATP, dTTP, dGTP, and dCTP). The four dye-labelled, chain-terminating dideoxy nucleotides are added as well, but in much smaller amounts than the ordinary nucleotides.

The mixture is first heated to denature the template DNA (separate the strands), then cooled so that the primer can bind to the single-stranded template. Once the primer has bound, the temperature is raised again, allowing DNA polymerase to synthesize new DNA starting from the primer. DNA polymerase will continue adding nucleotides to the chain until it happens to add a dideoxy nucleotide instead of a normal one. At that point, no further nucleotides can be added, so the strand will end with the dideoxy nucleotide.

NOTES

This process is repeated in a number of cycles. By the time the cycling is complete, it's virtually guaranteed that a dideoxy nucleotide will have been incorporated at every single position of the target DNA in at least one reaction. That is, the tube will contain fragments of different lengths, ending at each of the nucleotide positions in the original DNA (Refer Figure 13.5). The ends of the fragments will be labelled with dyes that indicate their final nucleotide.

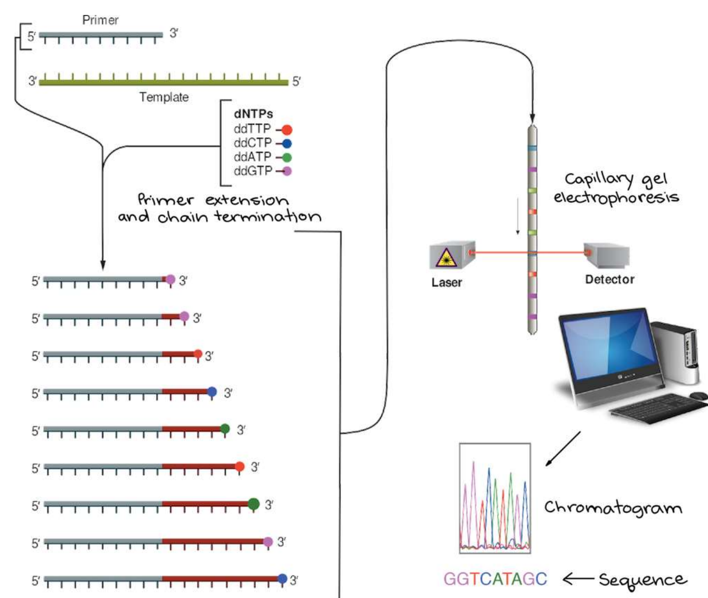


Fig. 13.4 *Sanger sequencing*

After the reaction is done, the fragments are run through a long, thin tube containing a gel matrix in a process called capillary gel electrophoresis. Short fragments move quickly through the pores of the gel, while long fragments move more slowly. As each fragment crosses the 'Finish Line' at the end of the tube, it's illuminated by a laser, allowing the attached dye to be detected.

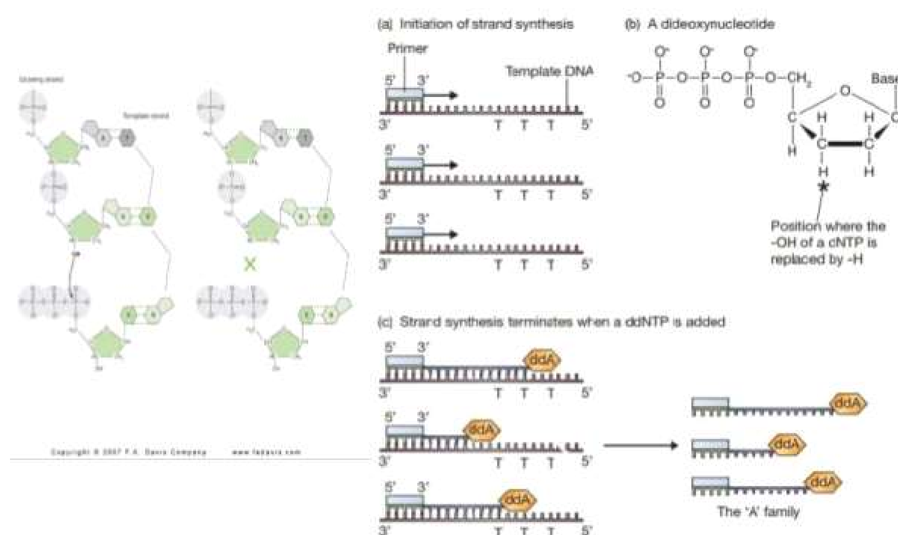


Fig. 13.5 *Sanger's Method*

The smallest fragment (ending just one nucleotide after the primer) crosses the finish line first, followed by the next-smallest fragment (ending two nucleotides after the primer), and so forth. Thus, from the colours of dyes registered one after another on the detector, the sequence of the original piece of DNA can be built up one nucleotide at a time. The data recorded by the detector consist of a series of peaks in fluorescence intensity, as shown in the chromatogram above. The DNA sequence is read from the peaks in the chromatogram.

Uses and Limitations

Sanger sequencing gives high-quality sequence for relatively long stretches of DNA (up to about 900900 base pairs). It's typically used to sequence individual pieces of DNA, such as bacterial plasmids or DNA copied in PCR.

However, Sanger sequencing is expensive and inefficient for larger-scale projects, such as the sequencing of an entire genome or metagenome (the 'Collective Genome' of a microbial community). For tasks, such as these, new, large-scale sequencing techniques are faster and less expensive.

Next-Generation Sequencing

The name may sound like Star Trek, but that's really what it's called! The most recent set of DNA sequencing technologies are collectively referred to as next-generation sequencing.

There are a variety of next-generation sequencing techniques that use different technologies. However, most share a common set of features that distinguish them from Sanger sequencing:

- Highly parallel: many sequencing reactions take place at the same time
- Micro scale: reactions are tiny and many can be done at once on a chip
- Fast: because reactions are done in parallel, results are ready much faster
- Low-cost: sequencing a genome is cheaper than with Sanger sequencing

13.3.4 Protein Synthesis in DNA

Proteins are giant molecules formed by polypeptide chains of hundreds to thousands of amino acids. These *polypeptide chains* are formed by about twenty kinds of *amino acids*. An amino acid consists of a basic amino group (-NH_2) and an acidic carboxyl group (-COOH). Different arrangement of amino acids in a polypeptide chain makes each protein unique. Proteins are fundamental constituents of protoplasm and building material of the cell.

They take part in the structural and functional organisation of the cell. Functional proteins like enzymes and hormones control the metabolism, biosynthesis, and energy production, and growth regulation, sensory and reproductive functions of the cell. Enzymes are catalysts in most of the biochemical reactions. Even the gene expression is controlled by enzymes. The replication of DNA and transcription of RNA is controlled by the **proteinous enzymes.**

NOTES

Components of Protein Synthesis

Protein synthesis is governed by the genetic information carried in the genes on DNA of the chromosomes.

The main components of the protein synthesis are:

- DNA
- Three Types of RNAs
- Amino Acids
- Ribosomes
- Enzymes.

DNA is the master molecule which possesses the genetic information about the sequence of amino acids in a polypeptide chain. Structure and properties of DNA regulate and control the synthesis of proteins. DNA present in the nucleus sends out information in the form of messenger RNA into the *Cytoplasm*, which is the site of the protein synthesis in eukaryotes. The messenger RNA carries the information regarding the sequence of amino acids of the polypeptide chain to be synthesized. This message or information is in the form of a *genetic code*. This genetic code specifies the language of amino acids to be assembled in a polypeptide. The genetic code is deciphered or translated into a sequence of amino acids.

Composition of Genetic Code

DNA molecule has three components. They are sugar, phosphates and nitrogen bases. Only nitrogen base sequence varies in different DNA molecules. Thus, the sequence of nitrogen bases or nucleotides in a DNA segment is the code or language in which the DNA sends out the message in the form of messenger RNA (mRNA).

The mRNA carries the genetic message (genetic code) in the form of nucleotide sequence. It has been found that there is co-linearity between nucleotide sequence of mRNA and amino acid sequence of the polypeptide chain synthesized.

The genetic code is the language of nitrogen bases. There are four kinds of nitrogen bases and twenty kinds of amino acids. Therefore four-letter language of nitrogen bases specifies the twenty letter language of amino acids.

13.4 INTRODUCTION AND CONCEPTS OF BIOLOGICAL DATABASE

Introduction

Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetic. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

Biological databases can be broadly classified into sequence, structure and functional databases. Nucleic acid and protein sequences are stored in sequence databases and structure databases store solved structures of RNA and proteins. Functional databases provide information on the physiological role of gene products, for example enzyme activities, mutant phenotypes, or biological pathways. Model Organism Databases are functional databases that provide species-specific data. Databases are important tools in assisting scientists to analyse and explain a host of biological phenomena from the structure of biomolecules and their interaction, to the whole metabolism of organisms and to understanding the evolution of species. By this knowledge helps facilitate the fight against diseases, assists in the development of medications, predicting certain genetic diseases and in discovering basic relationships among species in the history of life.

Biological knowledge is distributed among many different general and specialized databases. This sometimes makes it difficult to ensure the consistency of information. Integrative bioinformatics is one field attempting to tackle this problem by providing unified access. One solution is how biological databases cross-reference to other databases with accession numbers to link their related knowledge together.

Concepts

Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases. Biological database design, development, and long-term management is a core area of the discipline of bioinformatics. Data contents include gene sequences, textual descriptions, attributes and ontology classifications, citations, and tabular data. These are often described as semi-structured data, and can be represented as tables, key delimited records, and Extensible Mark-up Language (XML) structures. Figure 13.6 illustrate the Home Page of a Biological Database Called *STRING* which Characterises Functional Links between Proteins.

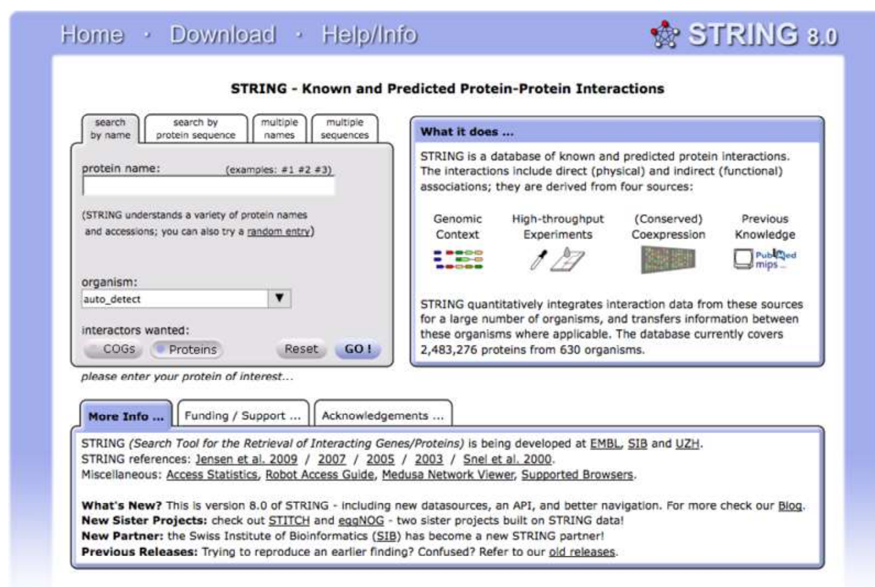


Fig.13.6 *Home Page of a Biological Database Called STRING*

NOTES

Nucleic Acids Research Database Issue

An important resource for finding biological databases is a special yearly issue of the journal Nucleic Acids Research (NAR). The Database Issue of NAR is freely available, and categorises many of the publicly available online databases related to biology and bioinformatics. A companion database to the issue called the ***Online Molecular Biology Database*** Collection lists 1,380 online databases. Other collections of databases exist, such as Meta Base and the Bioinformatics Links Collection.

Access

Mostly biological databases are available through the web sites that organise data such that users can browse through the data online. In addition the underlying data is usually available for download in a variety of formats. Biological Data comes in many formats. These formats include text, sequence data, protein structure and links. Each of these can be found from certain sources, for example:

- Text formats are provided by Primarily the MEDLINE Database (PubMed) and Online Mendelian Inheritance in Man (OMIM).
- Sequence data is provided by GenBank, in terms of DNA, and UniProt, in terms of protein.
- Protein structures are provided by Protein Data Bank (PDB), Structural Classification of Proteins (SCOP), and Protein Structure Classification database (CATH).

Biological Databases

- This databases consisting of biological data like protein sequencing, molecular structure, DNA sequences, etc., in an *organised form*.
- Several computer tools are there to manipulate the biological data like an update, delete, insert, etc., Scientists, researchers from all over the world enter their experiment data and results in a biological database so that it is available to a wider audience.
- Biological databases are free to use and contain a huge collection of a variety of biological data.

Uses of biological Databases

- It helps the researchers to study the available data and form a new thesis, anti-virus, helpful bacteria, medicines, etc.
- It helps scientists to understand the concepts of biological phenomena occurs in the world.
- The database acts as a storage of information.
- It helps remove the redundancy of data.

Types of Biological Databases

There are basically 3 types of biological databases are as follows.

1. Primary Databases
2. Secondary Database
3. Composite Databases

1. **Primary Databases:** It can also be called an archival database since it archives the experimental results submitted by the scientists. The primary database is populated with experimentally derived data like genome sequence, macromolecular structure, etc. The data entered here remains uncurated (no modifications are performed over the data). It obtains unique data obtained from the laboratory and these data are made accessible to normal users without any change. The data are given accession numbers when they are entered into the database. The same data can later be retrieved using the accession number. Accession number identifies each data uniquely and it never changes.

   Examples of Primary Databases

   - Examples of Primary database- Nucleic Acid Databases are Gen Bank and DNA Data Bank of Japan (DDBJ)
   - Protein Databases are Protein Data Bank (PDB), SwissProt, Protein Information Resource (PIR), TrEMBL, Metacyc, etc.

2. **Secondary Database:** The data stored in these types of databases are the analysed result of the primary database. Computational algorithms are applied to the primary database and meaningful and informative data is stored inside the secondary database. The data here are highly curated (processing the data before it is presented in the database). A secondary database is better and contains more valuable knowledge compared to the primary database.

   Examples of Secondary Databases

   Examples of Secondary databases are as follows.

   InterPro (protein families, motifs, and domains)

   UniProt Knowledgebase (sequence and functional information on proteins)

3. **Composite Databases:** The data entered in these types of databases are first compared and then filtered based on desired criteria. The initial data are taken from the primary database, and then they are merged together based on certain conditions. It helps in searching sequences rapidly. Composite Databases contain non-redundant data.

   Examples of Composite Databases

   Examples of Composite Databases are as follows.

   Composite Databases - Otto Warburg Java Library (OWL), Non-Redundant Database (NRD) and Swiss port +TREMBL

Check Your Progress

1. What is genome of an animal?
2. Why genetic diversity is important in animals?
3. Explain about the DNA.
4. What is polynucleotides in DNA?
5. Elaborate on the uses and limitations of Sanger sequencing.
6. Explain the components of protein synthesis.
7. Define the term biological database.
8. Give the uses of biological database.

NOTES

13.5 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. The genomes of almost all living creatures, both plants and animals, consist of Deoxyribonucleic Acid (DNA), the chemical chain that includes the genes that code for different proteins and the regulatory sequences that turn those genes on and off.
2. Genetic diversity is important because it helps maintain the health of a population, by including alleles that may be valuable in resisting diseases, pests and other stresses. If the environment changes, a population that has a higher variability of alleles will be better able to evolve to adapt to the new environment.
3. Deoxyribose Nucleic Acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying the genetic instructions used in the growth, development, functioning, are reproduction of all known living organisms and many viruses.
4. The two DNA strands are also known as polynucleotides as they are composed of simpler monomeric units called Nucleotides. Each nucleotide is composed of one of four nitrogen-containing nucleobases (Cytosine[C], Guanine [G], Adenine [A] or Thymine [T]), a sugar called Deoxyribose, and a Phosphate group.
5. Sanger sequencing gives high-quality sequence for relatively long stretches of DNA (up to about 900900 base pairs). It's typically used to sequence individual pieces of DNA, such as bacterial plasmids or DNA copied in PCR.

   However, Sanger sequencing is expensive and inefficient for larger-scale projects, such as the sequencing of an entire genome or metagenome (the 'Collective Genome' of a microbial community). For tasks, such as these, new, large-scale sequencing techniques are faster and less expensive.

NOTES

6. Components of Protein Synthesis

   Protein synthesis is governed by the genetic information carried in the genes on DNA of the chromosomes.

   The main components of the protein synthesis are:

   - DNA
   - Three Types of RNAs
   - Amino Acids
   - Ribosomes
   - Enzymes.

7. Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetic. Information contained in biological databases includes gene function, structure, localization (both cellular and chromosomal), clinical effects of mutations as well as similarities of biological sequences and structures.

8. Uses of biological Databases

   - It helps the researchers to study the available data and form a new thesis, anti-virus, helpful bacteria, medicines, etc.
   - It helps scientists to understand the concepts of biological phenomena occurs in the world.
   - The database acts as a storage of information.
   - It helps remove the redundancy of data.

13.6 SUMMARY

- Animals are highly diverse. Members of the animal kingdom are among the most conspicuous living things in the world. Animal evolution began in the ocean over 600 million years ago with tiny creatures that probably do not resemble any living organism today.
- The animal classification system characterizes animals based on their anatomy, morphology, evolutionary history, features of embryological development, and genetic makeup. This classification scheme is constantly developing as new information about species arises.
- Animals are the eaters or consumers of the earth. They are heterotrophs and depend directly or indirectly on plants, photosynthetic Protists (Algae), or autotrophic bacteria for nourishment. Animals are able to move from place to place in search of food. In most, ingestion of food is followed by digestion in an internal cavity.

- All animals are multicellular heterotrophs. The unicellular heterotrophic organisms called Protozoa, which were at one time regarded as simple animals, are now considered to be members of the kingdom Protista, the large and diverse group.
- The ability of animals to move more rapidly and in more complex ways than members of other kingdoms is perhaps their most striking characteristic and one that is directly related to the flexibility of their cells and the evolution of nerve and muscle tissues.
- Genes are the small segments present on DNA. They encode for proteins which describes an individual's phenotype. They are characterized by the presence of start codon in the beginning and stop codon at the end. There are few genes which do not code for functional proteins and many are hypothetical due to their unknown functions.
- The term Gene was coined by Johannsen in 1909 and he named hereditary units of Mendel as 'Genes'. Subsequently many concepts and views emerged on genes calling them as the hereditary component, thread like structures, etc.
- Gene delivery refers to the process of introducing foreign genetic material, such as Deoxyribonucleic Acid (DNA) or Ribonucleic Acid (RNA) directly into host cells.
- Deoxyribose Nucleic Acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying the genetic instructions used in the growth, development, functioning, are reproduction of all known living organisms and many viruses.
- DNA and Ribose Nucleic Acid (RNA) are Nucleic Acids; alongside Proteins, Lipids and complex Carbohydrates (Polysaccharides), Nucleic Acids are one of the four major types of macromolecules that are essential for all known forms of life.
- Nucleic acids were first isolated by Friedrich Miescher (1869) from pus cells. They were named nuclein. Hertwig (1884) proposed nuclein to be the carrier of hereditary traits. Because of their acidic nature they were named nucleinic acids and then nucleic acids (Altmann, 1899).
- The nitrogen bases of the two chains formed complementary pairs with purine of one and pyrimidine of the other held together by Hydrogen Bonds (A-T, C-G). Comple-mentary base pairing between the two polynucleotide chains is considered to be hall mark of their proposition.
- The two chains in DNA are twisted helically just as a rope ladder with rigid steps twisted into a spiral. Each turn of the spiral contains 10 nucleotides. This double helix or duplex model of DNA with antiparallel polynucleotide chains having complementary bases has an implicit mechanism of its replication and copying.

NOTES

- Linear DNA, without association with histone proteins, also occurs in some prokaryotes, for example Myco-plasma. In semi-autonomous cell organelles (mitochondria, plastids) DNA is circular, less commonly linear. It is always naked.
- DNA sequencing enables us to perform a thorough analysis of DNA because it provides us with the most basic information of all: the sequence of nucleotides. With this knowledge, for example we can locate regulatory and gene sequences, make comparisons between homologous genes across species and identify mutations.
- The DNA sample to be sequenced is combined in a tube with primer, DNA polymerase, and DNA nucleotides (dATP, dTTP, dGTP, and dCTP). The four dye-labelled, chain-terminating dideoxy nucleotides are added as well, but in much smaller amounts than the ordinary nucleotides.
- Sanger sequencing gives high-quality sequence for relatively long stretches of DNA (up to about 900900 base pairs). It's typically used to sequence individual pieces of DNA, such as bacterial plasmids or DNA copied in PCR.
- Proteins are giant molecules formed by polypeptide chains of hundreds to thousands of amino acids. These polypeptide chains are formed by about twenty kinds of amino acids.
- Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis.
- Biological databases can be broadly classified into sequence, structure and functional databases. Nucleic acid and protein sequences are stored in sequence databases and structure databases store solved structures of RNA and proteins.
- Relational database concepts of computer science and Information retrieval concepts of digital libraries are important for understanding biological databases.
- Mostly biological databases are available through the web sites that organise data such that users can browse through the data online. In addition the underlying data is usually available for download in a variety of formats. Biological Data comes in many formats. These formats include text, sequence data, protein structure and links.

13.7 KEY WORDS

- **Genome diversity:** Genome diversity is differ to the total number of genetic characteristics in the genetic makeup of a species, it ranges widely from the number of species to differences within species and can be attributed to the span of survival for a species.

NOTES

- **Sexual Reproduction:** Most animals reproduce sexually. Animal eggs, which are non-motile, are much larger than the small, usually flagellated sperm. In animals, cells formed in meiosis function directly as gametes.
- **Deoxyribose Nucleic Acid (DNA):** Deoxyribose Nucleic Acid (DNA) is a molecule composed of two chains that coil around each other to form a double helix carrying the genetic instructions used in the growth, development, functioning, are reproduction of all known living organisms and many viruses.
- **Polypeptide chains:** Proteins are giant molecules formed by polypeptide chains of hundreds to thousands of amino acids. These polypeptide chains are formed by about twenty kinds of amino acids.
- **Biological databases:** Biological databases are libraries of biological sciences, collected from scientific experiments, published literature, high-throughput experiment technology, and computational analysis. They contain information from research areas including genomics, proteomics, metabolomics, microarray gene expression, and phylogenetic.

13.8 SELF-ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Explain the animal genome diversity.
2. What is genetic diversity?
3. Define the term DNA.
4. Illustrate the structure of DNA.
5. Elaborate on the types of DNA.
6. Interpret the functions of DNA.
7. Comprehend the Sanger sequencing in DNA.
8. Explain about the protein synthesis in DNA.
9. State the concept of biological database.
10. What is primary biological database?
11. Elaborate on the composite biological database.
12. Distinguish between the primary and secondary biological database.

Long-Answer Questions

1. Discuss briefly about the animal genome diversity.
2. Analyse the DNA with structure and functions.
3. Describe the Sanger sequencing method in DNA. Give the uses and limitation.

4. Explain briefly about the protein synthesis in DNA. Give the appropriate examples.
5. Briefly explain about the introduction and concept of biological database with the help of its types.

NOTES

13.9 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.

UNIT 14 PHYLOGENETIC ANALYSIS

NOTES

Structure

14.0 INTRODUCTION

In biology, phylogenetics is a part of systematics that addresses the inference of the evolutionary history and relationships among or within groups of organisms, for example species or more inclusive taxa. These relationships are specifically hypothesized by phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology, often under a specified model of evolution of these traits. The result of such an analysis is a phylogeny, also known as a phylogenetic tree, a diagrammatic hypothesis of relationships that reflects the evolutionary history of a group of organisms. The informations and guidelines of a phylogenetic tree can be living taxa or fossils, and represent the 'End' or the 'Present', in an evolutionary lineage.

Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism. In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor.

PHYLogeny Inference Package (PHYLIP) is referred as a free computational phylogenetics package of programs for inferring evolutionary trees (phylogenies). It comprises of 35 portable programs, i.e., the source code is written in the programming language C. PHYLIP is a complete phylogenetic analysis package which was developed by Joseph Felsestein at University of Washington. PHYLIP is used to find the evolutionary relationships between different organisms. ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods.

In this unit, you will study about the phylogenetic analysis using PHYLIP and ClustalW.

NOTES

14.1 OBJECTIVES

After going through this unit, you will be able to:

- Understand the basics of phylogenetic analysis
- Explain about the PHYLIP and ClustalW
- Elaborate on the phylogenetic analysis method using PHYLIP and ClustalW

14.2 PHYLOGENETIC ANALYSIS

In biology, **phylogenetics** word is derived from two Greek words ***phylon***, which means tribe, clan, race, and ***genetikós***, which means origin, source, birth. Phylogenetics is a part of systematics that specifically addresses the inference of the evolutionary history and relationships among or within groups of organisms, for example species or more inclusive taxa. These relationships are typically hypothesized by means of phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology, often under a specified model of evolution of these traits. The result of such an analysis is termed as a phylogeny, also known as a phylogenetic tree, which is a diagrammatic hypothesis of relationships that reflects the evolutionary history of a group of organisms. The instructions and guidelines of a phylogenetic tree can be living taxa or fossils, and represent the 'End' or the 'Present' in an evolutionary lineage. A phylogenetic diagram can be rooted or unrooted. A rooted tree diagram indicates the hypothetical common ancestor or ancestral lineage of the tree while an unrooted tree diagram (a network) makes no assumption about the ancestral line, and therefore does not show the origin or 'Root' of the taxa being studied or the direction of inferred evolutionary transformations.

Additionally, the proper use for inferring phylogenetic patterns among taxa, phylogenetic analyses are often used to represent relationships among gene copies or individual organisms. Such uses are central to understand biodiversity, evolution, ecology, and genomes. Recently, in February 2021, scientists reported, for the first time, the sequencing of DNA from animal remains, a mammoth in this instance, over a million years old, the oldest DNA sequenced to date.

Taxonomy is the identification, naming and classification of organisms. Classifications are now usually based on phylogenetic data, and many systematists contend that only monophyletic taxa should be recognized as named groups. The degree to which classification depends on inferred evolutionary history differs depending on the school of taxonomy as follows:

- Phenetics ignores phylogenetic speculation altogether, instead try to represent the similarity between organisms.

- Cladistics (phylogenetic systematics) tries to reflect phylogeny in its classifications by only recognizing groups based on shared, derived characters (synapomorphies).
- Evolutionary taxonomy tries to take into account both the branching pattern and 'Degree of Difference' to find a compromise between them.

Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism.

Phylogenetic Tree

A phylogenetic tree, also termed as phylogeny or evolutionary tree, is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics. All life on Earth is part of a single phylogenetic tree, indicating common ancestry.

In a rooted phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of those descendants, and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit. Internal nodes are generally called hypothetical taxonomic units, as they cannot be directly observed. Trees are useful in fields of biology, such as bioinformatics, systematics, and phylogenetics. Unrooted trees illustrate only the relatedness of the leaf nodes and do not require the ancestral root to be known or inferred.

The idea of a 'Phylogenetic Tree' or the 'Tree of Life' arose from ancient notions of a ladder-like progression from lower into higher forms of life, such as in the 'Great Chain of Being'. Early representations of 'Branching' phylogenetic trees include a 'Paleontological Chart' showing the geological relationships among plants and animals in the book "***Elementary Geology***", by Edward Hitchcock (First Edition: 1840).

Charles Darwin (1859) also produced one of the first illustrations and crucially popularized the notion of an 'Evolutionary Tree' in his seminal book, "***The Origin of Species***". Over a century later, evolutionary biologists still use tree diagrams to depict evolution because such diagrams effectively convey the concept that speciation occurs through the adaptive and semi-random splitting of lineages. Over time, species classification has become less static and more dynamic.

Properties of Phylogenetic Tree

Following are the characteristic properties of the phylogenetic tree:

Rooted Phylogenetic Tree: A rooted phylogenetic tree is a directed tree with a unique node, the root, corresponding to the (usually attributed) most recent common ancestor of all the entities at the leaves of the tree. The root node does not have a parent node, but serves as the parent of all other nodes in the tree. The root is, therefore, a node of 'degree 2', while other internal nodes have a minimum degree of 3, where the term 'degree' refers to the total number of incoming and outgoing edges.

The most common method for rooting trees is the use of a definitive out-group, close enough to allow inference from trait data or molecular sequencing, but far enough to be a clear out-group.

Unrooted Phylogenetic Tree: Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. They do not require the ancestral root to be known or inferred. Unrooted trees can always be generated from rooted ones by simply omitting the root. By contrast, inferring the root of an unrooted tree requires some means of identifying ancestry. This is normally done by including an out-group in the input data so that the root is necessarily between the out-group and the rest of the taxa in the tree, or by introducing additional assumptions about the relative rates of evolution on each branch, such as an application of the molecular clock hypothesis.

Bifurcating versus Multifurcating Phylogenetic Tree: Both rooted and unrooted trees can be either bifurcating or multifurcating. A rooted bifurcating tree has exactly two descendants arising from each interior node, i.e., it forms a binary tree, and an unrooted bifurcating tree takes the form of an unrooted binary tree, a free tree with exactly three neighbours at each internal node. In contrast, a rooted multifurcating tree may have more than two children at some nodes and an unrooted multifurcating tree may have more than three neighbours at some nodes.

Labelled versus Unlabelled Phylogenetic Tree: Both rooted and unrooted trees can be either labelled or unlabelled. A labelled tree has specific values assigned to its leaves, while an unlabelled tree, sometimes called a tree shape, defines a topology only. Some sequence-based trees built from a small genomic locus, such as Phylotree, feature internal nodes labelled with inferred ancestral haplotypes.

Branching Diagrams in Phylogenetic Analysis

In **phylogenetic analysis**, **branching diagrams** are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor.

The diagram is known as a **phylogenetic tree**. Phylogenetic analysis is very significant for gathering collecting information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution.

With advancements in genetic sequencing techniques, phylogenetic analysis now includes the sequence of a gene to understand the evolutionary relationships among species. DNA being the hereditary material can now be sequenced easily, quickly, and cost-effectively, and the data thus obtained from genetic sequencing is very informative and explicit.

In a phylogenetic tree, leaves represent and specify the species, populations, individuals or genes which can be connected to nodes through branches (external branch). The branches represent the passage of genetic information between

NOTES

subsequent generations, and branch lengths denote genetic change or divergence. The degree of divergence is normally estimated by means of the average number of nucleotide substitutions per site.

While analysing a phylogenetic tree from the root toward the tips, a node represents the exact position from where two or more descendant lineages are generated from an ancestral lineage.

The specific or precise branching pattern typically created by lineage splitting is termed as topology. Topology represents or specifies the evolutionary development of the contemporary generation through progressive branching of the lineages.

A phylogenetic tree can be either rooted or unrooted, scaled or unscaled. The appropriate rooting of a phylogenetic tree is essential for enhanced understanding about the directionality of evolution and genetic divergence. In a rooted phylogenetic tree, various methods exist to accurately estimate the tree root using gene sequencing data and assumptions, such as the molecular clock, the midpoint rooting, out-group rooting, etc., while the unrooted phylogenetic tree simply represents the relationships among the species without actually illustrating an ancestral root of origin. In a scaled phylogenetic tree, a proportional relationship exists between the branch length and the amount of genetic divergence that happened on that branch. In an unscaled phylogenetic tree, all branches are of equal length and there is no correlation between the branch length and genetic divergence.

Figure 14.1 illustrates a phylogenetic analysis of Human Immunodeficiency Viruses or HIV (sequenced and compared) which is typically analysed by a Florida dentist, on the basis of data collected on his patients and other infected individuals from the local community. The phylogenetic tree is specifically represented as a circular diagram in which the most closely related taxa group or cluster together near the periphery while the deeper (older) associates are closer to the center of the circle. On the basis of the epidemiological studies of disease-causing organisms, it became possible to represent the analysed data in the form of phylogenetic tree because many viruses evolve at very rapid rates, often as high as several substitutions per thousand nucleotide sites per year. Figure 14.1 shows that the phylogenetic analysis is consistent with the suspected transmission of the virus from the dentist to six of his patients (those in the shaded dental clade). Four additional patients (D, H, F, and J) with risk factors for HIV, give the impression that they have been infected from other sources. One patient (J) typically give the impression that he/she might have been infected from two different sources. In Figure 14.1 the term 'LC' refers to Local Control.

NOTES

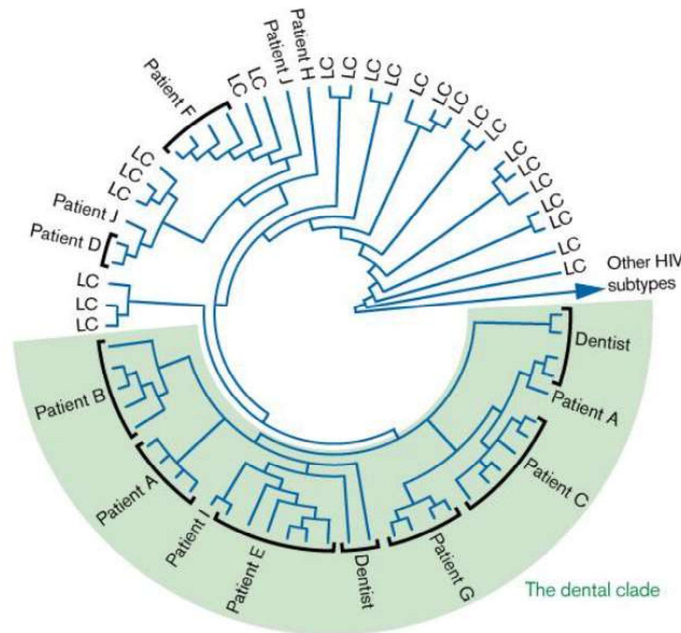


Fig. 14.1 *Phylogenetic Analysis of Human Immunodeficiency Viruses or HIV*

(Source: Wikipedia)

Optimality Criteria for Phylogenetic Analysis

The phylogenetic relationships can be analysed and estimated using the optimality criteria. To estimate the phylogenetic relationships of the taxa, a set of observations for a series of taxa is required. The method to measure and analyse that how the observed data can be best fit to alternative phylogenetic trees, requires an implicit or explicit model of evolution. For any phylogenetic analysis, for a given model of evolution and the observed data, three commonly used criteria are recommended for evaluating the fit of the data to trees, namely the parsimony, maximum likelihood and minimum evolution.

The first criterion is the **parsimony criterion** which is considered as the simplest form of criterion. To obtain the parsimony score, for each tree to be evaluated, the minimum possible number of changes for each ‘Character’, for example nucleotide position or morphological trait, is calculated and the minimum number of changes across all characters are evaluated and summed. The best phylogenetic tree is the one that requires the least or minimum changes across all characters.

The second criterion is the **maximum likelihood**, which is referred as the commonly used criterion. The best phylogenetic tree under maximum likelihood criterion is the one for which the observed data are the most feasible or probable for given an assumed model of evolution. Subsequently if the calculated and evaluated probabilities for any given tree are very low, then it is normally expected to take the log of the probability of the data so that the numbers can be easily

evaluated. Therefore, the maximum likelihood scores are negative numbers, and the best phylogenetic tree is the one with the log-likelihood closest to zero.

The third criterion is the **minimum evolution**, which holds characteristics of both of the above mentioned criteria, and is occasionally used for evaluating the fit of data to a phylogenetic tree. An explicit evolutionary model is typically used to 'Correct' observed differences, for example between all the pairs of the nucleotide or protein sequences being compared. Corrected evolutionary distances can be larger than the observed distances between the pairs of sequences, because they also account for superimposed changes, i.e., where a given nucleotide position has changed more than once since the two sequences diverged. To evaluate a given minimum evolution tree, the branch lengths on the tree are adjusted so that the path-length distances, i.e., the distance from one taxon to another along the tree, are as close as possible to the corrected distances (as evaluated using the least squares method). Once an optimal fit has been obtained for all the trees to be evaluated, then the best tree is selected as the tree with the lowest sum of branch lengths.

Enumerating Trees

The number of possible trees for a given number of leaf nodes depends on the specific type of tree, but there are always more labelled than unlabelled trees, more multifurcating than bifurcating trees, and more rooted than unrooted trees. The last distinction is the most biologically relevant; it arises because there are many places on an unrooted tree to put the root. For bifurcating labelled trees, the total number of rooted trees is calculated as:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!} \text{ for } n \geq 2,$$

Where n represents the number of leaf nodes.

For bifurcating labelled trees, the total number of unrooted trees is calculated as:

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!} \text{ for } n \geq 3.$$

Among labelled bifurcating trees, the number of unrooted trees with n leaves is equal to the number of rooted trees with $n - 1$ leaves.

The number of rooted trees grows quickly as a function of the number of tips. For 10 tips, there are more than 34×10^6 possible bifurcating trees, and the number of multifurcating trees increases faster.

Special Tree Types

Dendrogram: A dendrogram is a general name for a tree, whether phylogenetic or not, and hence also for the diagrammatic representation of a phylogenetic tree.

NOTES

Cladogram: A cladogram only represents a branching pattern; i.e., its branch lengths do not represent time or relative amount of character change, and its internal nodes do not represent ancestors.

Phylogram: A phylogram is a phylogenetic tree that has branch lengths proportional to the amount of character change. A chronogram is a phylogenetic tree that explicitly represents time through its branch lengths.

Dahlgrenogram: A dahlgrenogram is a diagram representing a cross section of a phylogenetic tree.

Phylogenetic Network: A phylogenetic network is not strictly speaking a tree, but rather a more general graph, or a directed acyclic graph in the case of rooted networks. They are specifically used to overcome some of the limitations inherent to trees.

Construction of Phylogenetic Tree

Phylogenetic trees composed with a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods, such as neighbor-joining or UPGMA (Unweighted Pair Group Method with Arithmetic mean), which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model. Many sequence alignment methods, such as ClustalW also create trees by using the simpler algorithms (i.e. those based on distance) of tree construction. Maximum parsimony is another simple method of estimating phylogenetic trees, but implies an implicit model of evolution (i.e., parsimony). More advanced methods use the optimality criterion of maximum likelihood, often within a Bayesian framework, and apply an explicit model of evolution to phylogenetic tree estimation. Identifying the optimal tree using many of these techniques is NP-hard, so heuristic search and optimization methods are used in combination with tree-scoring functions to identify a reasonably good tree that fits the data.

Tree building methods can be assessed on the basis of following criteria:

- Efficiency: How long does it take to compute the answer, how much memory does it need?
- Power: Does it make good use of the data or is information being wasted.
- Consistency: Will it converge on the same answer repeatedly, if each time given different data for the same model problem.
- Robustness: Does it cope well with violations of the assumptions of the underlying model.
- Falsifiability: Does it alert us when it is not good to use, i.e., when assumptions are violated?

Tree building techniques have also gained the attention of mathematicians. Trees can also be built using T-theory.

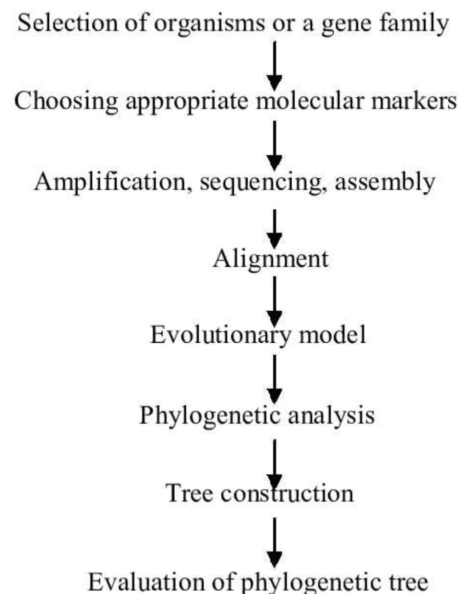
Steps in Phylogenetic Analysis

The basic steps in any phylogenetic analysis include the following:

1. **Assemble and Align a Dataset**
    - The first step is to identify a protein or DNA sequence of interest and assemble a dataset consisting of other related sequences.
    - DNA sequences of interest can be retrieved using NCBI BLAST or similar search tools.
    - Once sequences are selected and retrieved, multiple sequence alignment is created.
    - This involves arranging a set of sequences in a matrix to identify regions of homology.
    - There are many websites and software programs, such as ClustalW, MSA, MAFFT, and T-Coffee, designed to perform multiple sequence on a given set of molecular data.
2. **Build (Estimate) Phylogenetic Trees from Sequences using Computational Methods and Stochastic Models**
    - To build phylogenetic trees, statistical methods are applied to determine the tree topology and calculate the branch lengths that best describe the phylogenetic relationships of the aligned sequences in a dataset.
    - The most common computational methods applied include distance-matrix methods, and discrete data methods, such as maximum parsimony and maximum likelihood.
    - There are several software packages, such as Paup, PAML, PHYLIP, that apply these most popular methods.
3. **Statistically Test and Assess the Estimated Trees**
    - Tree estimating algorithms generate one or more optimal trees.
    - This set of possible trees is subjected to a series of statistical tests to evaluate whether one tree is better than another – and if the proposed phylogeny is reasonable.
    - Common methods for assessing trees include the Bootstrap and Jackknife Resampling methods, and analytical methods, such as parsimony, distance, and likelihood.

NOTES

Following steps are involved in the phylogenetic analysis:



Limitations of Phylogenetic Analysis

Although phylogenetic trees produced on the basis of sequenced genes or genomic data in different species can provide evolutionary insight, these analyses have important limitations. Most importantly, the trees that they generate are not necessarily correct – they do not necessarily accurately represent the evolutionary history of the included taxa. As with any scientific result, they are subject to falsification by further study (e.g., gathering of additional data, analysing the existing data with improved methods). The data on which they are based may be noisy; the analysis can be confounded by genetic recombination, horizontal gene transfer, hybridisation between species that were not nearest neighbours on the tree before hybridisation takes place, convergent evolution, and conserved sequences.

Also, there are problems in basing an analysis on a single type of character, such as a single gene or protein or only on morphological analysis, because such trees constructed from another unrelated data source often differ from the first, and therefore great care is needed in inferring phylogenetic relationships among species. This is most true of genetic material that is subject to lateral gene transfer and recombination, where different haplotype blocks can have different histories. In these types of analysis, the output tree of a phylogenetic analysis of a single gene is an estimate of the gene's phylogeny (i.e., a gene tree) and not the phylogeny of the taxa (i.e., species tree) from which these characters were sampled, though ideally, both should be very close. For this reason, thoughtful phylogenetic studies generally use a combination of genes that come from different genomic sources (e.g., from mitochondrial or plastid vs. nuclear genomes), or genes that would be

NOTES

expected to evolve under different selective regimes, so that homoplasy (false homology) would be unlikely to result from natural selection.

When extinct species are included as terminal nodes in an analysis (rather than, for example, to constrain internal nodes), they are considered not to represent direct ancestors of any extant species. Extinct species do not typically contain high-quality DNA.

The range of useful DNA materials has expanded with advances in extraction and sequencing technologies. Development of technologies able to infer sequences from smaller fragments, or from spatial patterns of DNA degradation products, would further expand the range of DNA considered useful.

Phylogenetic trees can also be inferred from a range of other data types, including morphology, the presence or absence of particular types of genes, insertion and deletion events – and any other observation thought to contain an evolutionary signal.

Phylogenetic networks are used when bifurcating trees are not suitable, due to these complications which suggest a more reticulate evolutionary history of the organisms sampled.

14.2.1 Phylogenetic Analysis using PHYLIP

PHYLIP (the **PHYLogeny Inference Package**) is a package of programs for inferring phylogenies (evolutionary trees). It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed. In particular, already-compiled executables are available for Windows (95/98/NT/2000/Me/XP/Vista), Mac OS X, and Linux systems. Complete documentation is available on the documentation files that come with the package.

Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences, gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

The programs are typically controlled through a menu, which queries the users which options they want to set, and allows them to start the computation. The data are read into the program from a text file, which the user can prepare using any word processor or text editor. It is important to note that this text file should not be in the special format of that word processor, i.e., it should instead be in 'Flat ASCII' or 'Text Only' format. Some sequence analysis programs, such as the **ClustalW alignment** program can write data files in the **PHYLIP format**. Most of the programs look for the data in a file called '**`infile`**', if they do not find this file they then ask the user to type in the file name of the data file.

Output is written onto special files with names like '**`outfile`**' and '**`outtree`**'. Trees written onto '**`outtree`**' are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.

PHYLIP is probably the most widely-distributed phylogeny package. PHYLIP is also the oldest widely-distributed package. It has been in distribution since October, 1980, and is still being updated.

Table 14.1 gives the list of programs that are included in PHYLIP.

Table 14.1 *Programs Listed in PHYLIP*

Program Name	Description
protpars	Estimates phylogenies of peptide sequences using the parsimony method
dnapars	Estimates phylogenies of DNA sequences using the parsimony method
dnapenny	DNA parsimony branch and bound method, finds all of the most parsimonious phylogenies for nucleic acid sequences by branch-and-bound search
dnamove	Interactive construction of phylogenies from nucleic acid sequences, with their evaluation by DNA parsimony method, with compatibility and display of reconstructed ancestral bases
dnacomp	Estimates phylogenies from nucleic acid sequence data using the compatibility criterion
dnaml	Estimates phylogenies from nucleotide sequences using the maximum likelihood method
dnamlk	DNA maximum likelihood method with molecular clock; using both dnaml and dnamlk together permits a likelihood-ratio test for the molecular clock hypothesis
proml	Estimates phylogenies from protein amino acid sequences by using the maximum likelihood method
promlk	Protein sequence maximum likelihood method with molecular clock
restml	Estimation of phylogenies by maximum likelihood using restriction sites data; not from restriction fragments but from the presence or absence of individual sites
dnainvar	For nucleic acid sequence data on four species, computes Lake's and Cavender's phylogenetic invariants, which test alternative tree topologies
dnadist	DNA distance method which computes four different distances between species from nucleic acid sequences; distances can then be used in the distance matrix programs
protdist	Protein sequence distance method which computes a distance measure for sequences, using maximum likelihood estimates based on the Dayhoff PAM matrix, Kimura's 1983 approximation to it, or a model based on genetic code plus a constraint on changing to a different category of amino acid
restdist	Distances calculated from restriction sites data or restriction fragments data

NOTES

seqboot	Bootstrapping-Jackknifing program; reads in a data set, and emits multiple data sets from it by bootstrap resampling
fitch	Fitch-Margoliash distance matrix method; estimates phylogenies from distance matrix data under the *additive tree model* according to which the distances are expected to equal the sums of branch lengths between species
kitsch	Fitch-Margoliash distance matrix method with molecular clock; estimates phylogenies from distance matrix data under the *ultrametric* model which is the same as the additive tree model except an evolutionary clock is assumed
neighbor	Implementation of the methods neighbor joining and UPGMA
contml	Maximum likelihood continuous characters and gene frequencies; estimates phylogenies from gene frequency data by maximum likelihood under a model in which all divergence is due to genetic drift in the absence of new mutations; also does maximum likelihood analysis of continuous characters that evolve by a Brownian Motion model, assuming that the characters evolve at equal rates and in an uncorrelated fashion; does not account for character correlations
contrast	Reads a tree from a tree file, and a data set with continuous characters data, and emits the independent contrasts for those characters, for use in any multivariate statistics package
gendist	Genetic distance program which computes one of three different genetic distance formulas from gene frequency data
pars	Unordered multistate discrete-characters parsimony method
mix	Estimates phylogenies by some parsimony methods for discrete character data with two states (0, 1); allows using methods: Wagner, Camin-Sokal, or arbitrary mixes
penny	Branch and bound mixed method which finds all of the most parsimonious phylogenies for discrete-character data with two states, for the Wagner, Camin-Sokal, and mixed parsimony criteria using the branch-and-bound method of exact search
move	Interactive construction of phylogenies from discrete character data with two states (0, 1); evaluates parsimony and compatibility criteria for those phylogenies and displays reconstructed states throughout the tree
dollop	Estimates phylogenies by the Dollo or polymorphism parsimony criteria for discrete character data with two states (0, 1)
dolpenny	Finds all or most parsimonious phylogenies for discrete-character data with two states, for the Dollo or polymorphism parsimony criteria using the branch-and-bound method of exact search
dolmove	Interactive construction of phylogenies from discrete character data with two states (0, 1) using the Dollo or polymorphism parsimony criteria; evaluates parsimony and compatibility criteria for those phylogenies; displays reconstructed states throughout the tree

clique	Finds the largest clique of mutually compatible characters, and the phylogeny which they recommend, for discrete character data with two states (0, 1); the largest clique (or all cliques within a given size range of the largest one) are found by a fast branch and bound search method
factor	Character recoding program which takes discrete multistate data with character state trees and emits the corresponding data set with the two states (0, 1)
drawgram	Rooted tree drawing program which plots rooted phylogenies, cladograms, and phenograms in a wide variety of user-controllable formats. The program is interactive and allows previewing of the tree on PC or Macintosh graphics screens, and Tektronix or Digital graphics terminals.
drawtree	Unrooted tree drawing program similar to DRAWGRAM, but plots phylogenies
consense	Consensus tree program which computes trees by the majority-rule tree method, which also allows easily finding the strict consensus tree; unable to compute Adams consensus tree
treedist	Computes the Robinson–Foulds symmetric difference distance between trees, which allows differences in tree topology
retree	Interactive tree rearrangement program which reads in a tree (with branch lengths if needed) and allows rerooting the tree, to flip branches, to change species names and branch lengths, and then write the result out; can be used to convert between rooted and unrooted trees

1. Phylogenetic Analysis using PHYLIP - Rooted Trees

Procedure for phylogenetic analysis includes the following steps.

First go to simulator tab to know more about how to retrieve the query sequence.

Using PHYLIP

Align the multiple DNA sequences (output of the ClustalW) and save it in PHYLIP format as infile.phy.

Start the program of **dnadist** by clicking the icon and giving this infile as input.

All the PHYLIP programs are menu driven programs.

The dnadist will calculate pairwise distances between the sequences.

Initially, the dnadist will ask whether the input file is there in the PHYLIP folder. If the file does not exist, it will ask you to give the correct file name.

After giving the correct input, if required, it will ask to change any settings for the program by typing the first letter or number.

If the changes are not required, by typing 'Y' it will start running the program.

Output will return to the file as 'outfile', so that the output of this file can be used as input of another program. Following program illustrates the distance representation.

NOTES

```
  0           Terminal type (IBM PC, ANSI, none)?  IBM PC
  1            Print out the data at start of run  No
  2          Print indications of progress of run  Yes

  Y to accept these or type the letter for one to change
y
Distances calculated for species
    HUMAN         .....
    CHIMPANZEP    ....
    MONKEYPAPo    ...
    MONKEYMOCA    ..
    RABBIT1       .
    RABBIT2

Distances written to file "outfile"

Done.

Press enter to quit.
```

Output would be as represented as shown below.

```
    6
HUMAN       0.000000 0.035305 0.049562 0.040798 0.222263 0.223706
CHIMPANZEP  0.035305 0.000000 0.074717 0.042751 0.222159 0.223601
MONKEYPAPo  0.049562 0.074717 0.000000 0.036918 0.222306 0.221246
MONKEYMOCA  0.040798 0.042751 0.036918 0.000000 0.222165 0.223678
RABBIT1     0.222263 0.222159 0.222306 0.222165 0.000000 0.014499
RABBIT2     0.223706 0.223601 0.221246 0.223678 0.014499 0.000000
```

Like dnadist, neighbor also gives sequence distance analysis. Output of dnadist is given as input to neighbor. The output file and tree file will be returned to 'outfile' and 'outtree' as represented below. It represents the sequence distance analysis.

```
  0    Terminal type (IBM PC, ANSI, none)?  IBM PC
  1     Print out the data at start of run  No
  2   Print indications of progress of run  Yes
  3                         Print out tree  Yes
  4         Write out trees onto tree file?  Yes

  Y to accept these or type the letter for one to change
y
Cycle    3: species 5 (    0.00683) joins species 6 (    0.00767)
Cycle    2: species 1 (    0.01315) joins species 2 (    0.02215)
Cycle    1: node 1 (    0.00862) joins node 5 (    0.18941)
last cycle:
 node 1  (    0.00723) joins species 3  (    0.02326) joins species 4  (    0.01365
)

Output written on file "outfile"

Tree written on file "outtree"

Done.

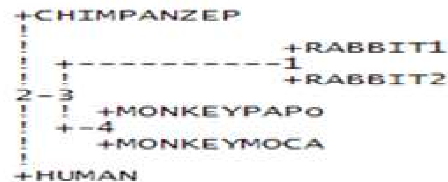
Press enter to quit.
```

Branch lengths and tree are represented with the help of neighbor joining method. The 'outfile' and 'outtree' after the neighbor joining method are given below. Following image shows an 'outfile'.

NOTES

```
   6 Populations
Neighbor-Joining/UPGMA method version 3.69

 Neighbor-joining method
 Negative branch lengths allowed
```



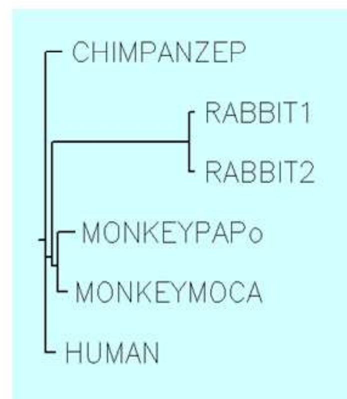

```
remember: this is an unrooted tree!
```

Between	And	Length
2	CHIMPANZEP	0.02215
2	3	0.00862
3	1	0.18941
1	RABBIT1	0.00683
1	RABBIT2	0.00767
3	4	0.00723
4	MONKEYPAPO	0.02326
4	MONKEYMOCA	0.01365
2	HUMAN	0.01315

Following image shows an 'outtree'.

```
(CHIMPANZEP:0.02215,((RABBIT1:0.00683,RABBIT2:0.00767):0.18941,
(MONKEYPAPo:0.02326,MONKEYMOCA:0.01365):0.00723):0.00862,HUMAN:0.01315);
```

Rooted trees are represented by means of **drawgram** by providing the 'input' as the 'outfile' obtained from **neighbor joining** method. Rooted tree specify and consider an imaginary root as the start from which the other sequences are associated and aligned. Following image shows the **rooted tree**.



2. Phylogenetic Analysis using PHYLIP - Unrooted Trees

PHYLIP File Format

Following are the significant points that specify the PHYLIP file format.

- The input files have information about the number of sequences, nucleic acids and amino acids.

NOTES

- The sequence has 10 characters length. Spaces can be added to the end of the short sequences to make them long.
- Gaps can be represented as '-'.
- Missing data can be represented as '?'.
- Spaces between the alignments are allowed usually after every 10 bases.

Example for 4 1061

```
GGCCTGCTCT  GCCTG- - - - -   CCCTGGCTTC AAGAGGG—C   AGTGCCTTCC
AGACGGAAAA AAAGGAAAAG   TCACGACATC CCCAA- - - - C   AGCCCCTCCA
- - - - - - - - -   - - - ? - - - - -G   CCGTGGTA- -  - - - - - - - - - -   - - - - GATTTG
```

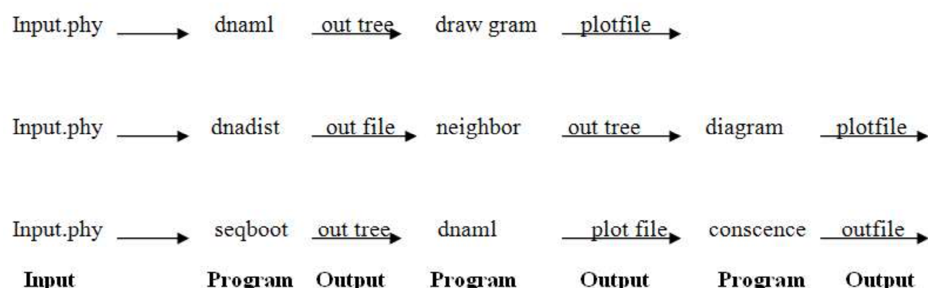
4 Indicates Number of Species Taken for Phylogenetic Analysis.

1061 Indicates Number of Characters.

PHYLIP Program

The PHYLIP programs run in sequential manner and the output of one program is used as input of another program. User must know how to use these programs in a sequential manner. Simple examples for running the PHYLIP programs are specified in the below given flowchart.

Flowchart



14.2.2 Phylogenetic Analysis using ClustalW

Clustal is a series of widely used computer programs typically used in bioinformatics for multiple sequence alignment. There are many different versions of Clustal over the development of the algorithm. There have been many variations of the Clustal software, all of which are listed below:

- **Clustal:** The original software for multiple sequence alignments, created by Des Higgins in 1988, was based on deriving phylogenetic trees from pairwise sequences of amino acids or nucleotides.
- **ClustalV:** The second generation of the Clustal software was released in 1992 and was a rewrite of the original Clustal package. It introduced

phylogenetic tree reconstruction on the final alignment, the ability to create alignments from existing alignments, and the option to create trees from alignments using a method called Neighbor joining.

- **ClustalW:** The third generation, released in 1994, greatly improved upon the previous versions. It improved upon the progressive alignment algorithm in various ways, including allowing individual sequences to be weighted down or up according to similarity or divergence respectively in a partial alignment. It also included the ability to run the program in batch mode from the command line.
- **ClustalX:** This version, released in 1997, was the first to have a Graphical User Interface (GUI).
- **ClustalΩ (Omega):** The current standard version of Clustal.
- **Clustal2:** The updated versions of both **ClustalW** and **ClustalX** with higher accuracy and efficiency.

The additional recent versions of the Clustal software is available for Windows, Mac OS, and UNIX/Linux.

ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods. In these, the most similar sequences, that is, those with the best alignment score are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained. All variations of the Clustal software align sequences using a heuristic that progressively builds a multiple sequence alignment from a series of pairwise alignments.

Essentially, Clustal creates multiple sequence alignments through the following three main steps:

1. Do a pairwise alignment using the progressive alignment method.
2. Create a guide tree or use a user-defined tree.
3. Use the guide tree to carry out a multiple alignment.

These steps are carried out automatically when the 'Do Complete Alignment' option is selected. Other options include 'Do Alignment from Guide Tree and Phylogeny' and 'Produce Guide Tree Only'.

Input/Output

This program accepts a wide range of input formats, including NBRF/PIR, FASTA, EMBL/Swiss-Prot, Clustal, GCC/MSF, GCG9 RSF and GDE.

The output format can be one or many of the following:

Clustal, NBRF/PIR, GCG/MSF, PHYLIP, GDE or NEXUS.

NOTES

Table 14.2 illustrates the reading multiple sequence alignment output.

Table 14.2 *Reading Multiple Sequence Alignment Output*

Symbol	Definition	Meaning
*	Asterisk	Positions that have a single and fully conserved residue
:	Colon	Conservation between groups of strongly similar properties with a score greater than 0.5 on the PAM 250 matrix
.	Period	Conservation between groups of weakly similar properties with a score less than or equal to 0.5 on the PAM 250 matrix

The same symbols are shown for both DNA/RNA alignments and protein alignments, so while * (asterisk) symbols are useful to both, the other consensus symbols should be ignored for DNA/RNA alignments.

Settings: Many settings can be modified to adapt the alignment algorithm to different circumstances. The main parameters are the gap opening penalty and the gap extension penalty.

ClustalW Software Algorithm

Figure 14.2 depicts the steps of the ClustalW software algorithm specifically used for global alignments.

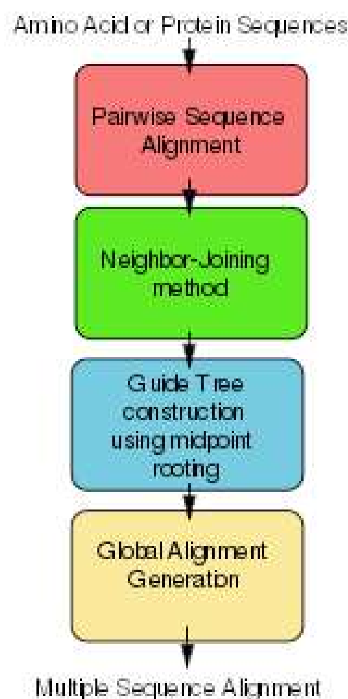


Fig. 14.2 *Steps of the ClustalW Software Algorithm for Global Alignments*

ClustalW like the other Clustal tools is used for aligning multiple nucleotide or protein sequences in an efficient manner. It uses progressive alignment methods, which align the most similar sequences first and work down to the least similar sequences until a global alignment is created. ClustalW is a matrix-based algorithm,

whereas tools like T-Coffee and Dialign are consistency-based. ClustalW has a fairly efficient algorithm that competes well against other software.

NOTES

Algorithm

ClustalW uses progressive alignment methods. In these, the sequences with the best alignment score are aligned first, then progressively more distant groups of sequences are aligned. This heuristic approach is essential due to the time and memory demand of finding the global optimal solution. The first step to the algorithm is computing a rough distance matrix between each pair of sequences, also known as pairwise sequence alignment. The next step is a neighbor-joining method that uses midpoint rooting to create an overall guide tree. The guide tree is then used as a rough template to generate a global alignment.

Time Complexity

ClustalW has a time complexity of $O(N^2)$ because of its use of the neighbor-joining method. In the updated version (ClustalW2) there is an option built into the software to use UPGMA which is faster with large input sizes. The command line flag in order to use it instead of neighbor-joining is:

```
-clustering=UPGMA
```

For example, on a standard desktop, running UPGMA on 10,000 sequences would produce results in less than a minute while neighbor-joining would take over an hour. By running the ClustalW algorithm with this adjustment, it saves significant amounts of time. ClustalW2 also has an option to use iterative alignment to increase alignment accuracy. Following are the various command line flags to achieve this:

```
-Iteration=Alignment
-Iteration=Tree
-numiters
```

The first command line option refines the final alignment. The second option incorporates the scheme into the progressive alignment step of the algorithm. The third specifies the number of iteration cycles where the default value is set to 3.

Accuracy and Results

The algorithm ClustalW provides a close-to-optimal result almost every time. However, it does exceptionally well when the data set contains sequences with varied degrees of divergence. This is because in a data set like this, the guide tree becomes less sensitive to noise. ClustalW was one of the first algorithms to combine pairwise alignment and global alignment in an attempt to be speed efficient.

ClustalW, when compared to other MSA (Multiple Sequence Alignment) algorithms, performed as one of the quickest while still maintaining a level of accuracy. The accuracy for ClustalW when tested against MAFFT, T-Coffee, Clustal Omega, and other MSA implementations had the lowest accuracy for full-length sequences. It had the least RAM (Random Access Memory) memory demanding algorithm out of all the ones tested in the study. While ClustalW recorded

the lowest level of accuracy among its competitors, it still maintained what some would estimate acceptable. There have been updates and improvements to the algorithm that are present in ClustalW2 that work to increase accuracy while still maintaining its greatly valued speed.

Check Your Progress

1. What is phylogenetics?
2. Elaborate on the degrees to which classification depends.
3. Define the idea of a 'Phylogenetic Tree' or the 'Tree of Life'.
4. In phylogenetic analysis, why branching diagrams are made?
5. For bifurcating labelled trees, how the total number of rooted trees and unrooted tree is calculated?
6. What is PHYLIP?

14.3 ANSWERS TO CHECK YOUR PROGRESS QUESTIONS

1. In biology, phylogenetics word is derived from two Greek words *phylon*, which means tribe, clan, race, and *genetikós*, which means origin, source, birth. Phylogenetics is a part of systematics that specifically addresses the inference of the evolutionary history and relationships among or within groups of organisms, for example species or more inclusive taxa. These relationships are typically hypothesized by means of phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology, often under a specified model of evolution of these traits.
2. The degree to which classification depends on inferred evolutionary history differs depending on the school of taxonomy as follows:
   - Phenetics ignores phylogenetic speculation altogether, instead try to represent the similarity between organisms.
   - Cladistics (phylogenetic systematics) tries to reflect phylogeny in its classifications by only recognizing groups based on shared, derived characters (synapomorphies).
   - Evolutionary taxonomy tries to take into account both the branching pattern and 'Degree of Difference' to find a compromise between them.

   Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism.
3. A phylogenetic tree, also termed as phylogeny or evolutionary tree, is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and

NOTES

differences in their physical or genetic characteristics. All life on Earth is part of a single phylogenetic tree, indicating common ancestry. The idea of a 'Phylogenetic Tree' or the 'Tree of Life' arose from ancient notions of a ladder-like progression from lower into higher forms of life, such as in the 'Great Chain of Being'. Early representations of 'Branching' phylogenetic trees include a 'Paleontological Chart' showing the geological relationships among plants and animals in the book "*Elementary Geology*", by Edward Hitchcock (First Edition: 1840).

4. In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor. The diagram is known as a phylogenetic tree.

   In a phylogenetic tree, leaves represent and specify the species, populations, individuals or genes which can be connected to nodes through branches (external branch). The branches represent the passage of genetic information between subsequent generations, and branch lengths denote genetic change or divergence. The degree of divergence is normally estimated by means of the average number of nucleotide substitutions per site.

   While analysing a phylogenetic tree from the root toward the tips, a node represents the exact position from where two or more descendant lineages are generated from an ancestral lineage.

5. For bifurcating labelled trees, the total number of rooted trees is calculated as:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!} \text{ for } n \geq 2,$$

   Where n represents the number of leaf nodes.

   For bifurcating labelled trees, the total number of unrooted trees is calculated as:

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!} \text{ for } n \geq 3.$$

   Among labelled bifurcating trees, the number of unrooted trees with n leaves is equal to the number of rooted trees with $n-1$ leaves.

6. PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). It is available free over the Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed.

NOTES

The programs are typically controlled through a menu, which queries the users which options they want to set, and allows them to start the computation. The data are read into the program from a text file, which the user can prepare using any word processor or text editor. It is important to note that this text file should not be in the special format of that word processor, i.e., it should instead be in 'Flat ASCII' or 'Text Only' format. Some sequence analysis programs, such as the ClustalW alignment program can write data files in the PHYLIP format. Most of the programs look for the data in a file called '`infile`', if they do not find this file they then ask the user to type in the file name of the data file.

14.4 SUMMARY

- In biology, phylogenetics word is derived from two Greek words *phylon*, which means tribe, clan, race, and *genetikós*, which means origin, source, birth. Phylogenetics is a part of systematics that specifically addresses the inference of the evolutionary history and relationships among or within groups of organisms, for example species or more inclusive taxa.
- These relationships are typically hypothesized by means of phylogenetic inference methods that evaluate observed heritable traits, such as DNA sequences or morphology, often under a specified model of evolution of these traits.
- The result of such an analysis is termed as a phylogeny, also known as a phylogenetic tree, which is a diagrammatic hypothesis of relationships that reflects the evolutionary history of a group of organisms.
- Taxonomy is the identification, naming and classification of organisms. Classifications are now usually based on phylogenetic data, and many systematists contend that only monophyletic taxa should be recognized as named groups.
- Phenetics ignores phylogenetic speculation altogether, instead try to represent the similarity between organisms.
- Cladistics (phylogenetic systematics) tries to reflect phylogeny in its classifications by only recognizing groups based on shared, derived characters (synapomorphies).
- Evolutionary taxonomy tries to take into account both the branching pattern and 'Degree of Difference' to find a compromise between them.
- Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism.
- A phylogenetic tree, also termed as phylogeny or evolutionary tree, is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and

NOTES

differences in their physical or genetic characteristics. All life on Earth is part of a single phylogenetic tree, indicating common ancestry.

- In a rooted phylogenetic tree, each node with descendants represents the inferred most recent common ancestor of those descendants, and the edge lengths in some trees may be interpreted as time estimates. Each node is called a taxonomic unit.
- Internal nodes are generally called hypothetical taxonomic units, as they cannot be directly observed. Trees are useful in fields of biology, such as bioinformatics, systematics, and phylogenetics. Unrooted trees illustrate only the relatedness of the leaf nodes and do not require the ancestral root to be known or inferred.
- The idea of a 'Phylogenetic Tree' or the 'Tree of Life' arose from ancient notions of a ladder-like progression from lower into higher forms of life, such as in the 'Great Chain of Being'. Early representations of 'Branching' phylogenetic trees include a 'Paleontological Chart' showing the geological relationships among plants and animals in the book "*Elementary Geology*", by Edward Hitchcock (First Edition: 1840).
- A rooted phylogenetic tree is a directed tree with a unique node, the root, corresponding to the (usually attributed) most recent common ancestor of all the entities at the leaves of the tree.
- The root node does not have a parent node, but serves as the parent of all other nodes in the tree. The root is, therefore, a node of 'degree 2', while other internal nodes have a minimum degree of 3, where the term 'degree' refers to the total number of incoming and outgoing edges.
- Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. They do not require the ancestral root to be known or inferred.
- Unrooted trees can always be generated from rooted ones by simply omitting the root. By contrast, inferring the root of an unrooted tree requires some means of identifying ancestry.
- Both rooted and unrooted trees can be either bifurcating or multifurcating. A rooted bifurcating tree has exactly two descendants arising from each interior node, i.e., it forms a binary tree, and an unrooted bifurcating tree takes the form of an unrooted binary tree, a free tree with exactly three neighbours at each internal node.
- A rooted multifurcating tree may have more than two children at some nodes and an unrooted multifurcating tree may have more than three neighbours at some nodes.
- Both rooted and unrooted trees can be either labelled or unlabelled. A labelled tree has specific values assigned to its leaves, while an unlabelled tree, sometimes called a tree shape, defines a topology only. Some sequence-

NOTES

based trees built from a small genomic locus, such as Phylotree, feature internal nodes labelled with inferred ancestral haplotypes.

- In phylogenetic analysis, branching diagrams are made to represent the evolutionary history or relationship between different species, organisms, or characteristics of an organism (genes, proteins, organs, etc.) that are developed from a common ancestor.
- The diagram is known as a phylogenetic tree. Phylogenetic analysis is very significant for gathering collecting information on biological diversity, genetic classifications, as well as learning developmental events that occur during evolution.
- In a phylogenetic tree, leaves represent and specify the species, populations, individuals or genes which can be connected to nodes through branches (external branch).
- The branches represent the passage of genetic information between subsequent generations, and branch lengths denote genetic change or divergence. The degree of divergence is normally estimated by means of the average number of nucleotide substitutions per site.
- While analysing a phylogenetic tree from the root toward the tips, a node represents the exact position from where two or more descendant lineages are generated from an ancestral lineage.
- The first criterion is the parsimony criterion which is considered as the simplest form of criterion. To obtain the parsimony score, for each tree to be evaluated, the minimum possible number of changes for each 'Character', for example nucleotide position or morphological trait, is calculated and the minimum number of changes across all characters are evaluated and summed.
- The best phylogenetic tree is the one that requires the least or minimum changes across all characters.
- The second criterion is the maximum likelihood, which is referred as the commonly used criterion. The best phylogenetic tree under maximum likelihood criterion is the one for which the observed data are the most feasible or probable for given an assumed model of evolution.
- The third criterion is the minimum evolution, which holds characteristics of both of the above mentioned criteria, and is occasionally used for evaluating the fit of data to a phylogenetic tree.
- An explicit evolutionary model is typically used to 'Correct' observed differences, for example between all the pairs of the nucleotide or protein sequences being compared.
- Corrected evolutionary distances can be larger than the observed distances between the pairs of sequences, because they also account for superimposed

changes, i.e., where a given nucleotide position has changed more than once since the two sequences diverged.

- Once an optimal fit has been obtained for all the trees to be evaluated, then the best tree is selected as the tree with the lowest sum of branch lengths.
- For bifurcating labelled trees, the total number of rooted trees is calculated as:

$$(2n-3)!! = \frac{(2n-3)!}{2^{n-2}(n-2)!} \text{ for } n \geq 2,$$

Where n represents the number of leaf nodes.

- For bifurcating labelled trees, the total number of unrooted trees is calculated as:

$$(2n-5)!! = \frac{(2n-5)!}{2^{n-3}(n-3)!} \text{ for } n \geq 3.$$

- Among labelled bifurcating trees, the number of unrooted trees with n leaves is equal to the number of rooted trees with $n-1$ leaves.
- A dendrogram is a general name for a tree, whether phylogenetic or not, and hence also for the diagrammatic representation of a phylogenetic tree.
- A cladogram only represents a branching pattern; i.e., its branch lengths do not represent time or relative amount of character change, and its internal nodes do not represent ancestors.
- A phylogram is a phylogenetic tree that has branch lengths proportional to the amount of character change. A chronogram is a phylogenetic tree that explicitly represents time through its branch lengths.
- A dahlgrenogram is a diagram representing a cross section of a phylogenetic tree.
- A phylogenetic network is not strictly speaking a tree, but rather a more general graph, or a directed acyclic graph in the case of rooted networks. They are specifically used to overcome some of the limitations inherent to trees.
- Phylogenetic trees composed with a nontrivial number of input sequences are constructed using computational phylogenetics methods. Distance-matrix methods, such as neighbor-joining or UPGMA (Unweighted Pair Group Method with Arithmetic mean), which calculate genetic distance from multiple sequence alignments, are simplest to implement, but do not invoke an evolutionary model.
- PHYLIP (the PHYLogeny Inference Package) is a package of programs for inferring phylogenies (evolutionary trees). It is available free over the

NOTES

Internet, and written to work on as many different kinds of computer systems as possible. The source code is distributed (in C), and executables are also distributed.

- In particular, already-compiled executables are available for Windows (95/98/NT/2000/Me/XP/Vista), Mac OS X, and Linux systems. Complete documentation is available on the documentation files that come with the package.
- Output is written onto special files with names like '`outfile`' and '`outtree`'. Trees written onto '`outtree`' are in the Newick format, an informal standard agreed to in 1986 by authors of a number of major phylogeny packages.
- Clustal is a series of widely used computer programs typically used in bioinformatics for multiple sequence alignment. There are many different versions of Clustal over the development of the algorithm.
- ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods. In these, the most similar sequences, that is, those with the best alignment score are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained.
- All variations of the Clustal software align sequences using a heuristic that progressively builds a multiple sequence alignment from a series of pairwise alignments.

14.5 KEY WORDS

- **Phylogenetics:** In biology, phylogenetics word is derived from two Greek words *phylon*, which means tribe, clan, race, and *genetikós*, which means origin, source, birth. Phylogenetics is a part of systematics that specifically addresses the inference of the evolutionary history and relationships among or within groups of organisms, for example species or more inclusive taxa.
- **Phylogenetic analysis:** Phylogenetic analysis is the study of the evolutionary development of a species or a group of organisms or a particular characteristic of an organism.
- **Phylogenetic tree:** A phylogenetic tree, also termed as phylogeny or evolutionary tree, is a branching diagram or a tree showing the evolutionary relationships among various biological species or other entities based upon similarities and differences in their physical or genetic characteristics.
- **Rooted phylogenetic tree:** A rooted phylogenetic tree is a directed tree with a unique node, the root, corresponding to the (usually attributed) most

recent common ancestor of all the entities at the leaves of the tree. The root node does not have a parent node, but serves as the parent of all other nodes in the tree.

- **Unrooted phylogenetic tree:** Unrooted trees illustrate the relatedness of the leaf nodes without making assumptions about ancestry. They do not require the ancestral root to be known or inferred. Unrooted trees can always be generated from rooted ones by simply omitting the root.
- **Dendrogram:** A dendrogram is a general name for a tree, whether phylogenetic or not, and hence also for the diagrammatic representation of a phylogenetic tree.
- **Cladogram:** A cladogram only represents a branching pattern; i.e., its branch lengths do not represent time or relative amount of character change, and its internal nodes do not represent ancestors.
- **Phylogram:** A phylogram is a phylogenetic tree that has branch lengths proportional to the amount of character change. A chronogram is a phylogenetic tree that explicitly represents time through its branch lengths.
- **Dahlgrenogram:** A dahlgrenogram is a diagram representing a cross section of a phylogenetic tree.
- **Phylogenetic network:** A phylogenetic network is not strictly speaking a tree, but rather a more general graph, or a directed acyclic graph in the case of rooted networks. They are specifically used to overcome some of the limitations inherent to trees.

14.6 SELF ASSESSMENT QUESTIONS AND EXERCISES

Short-Answer Questions

1. Define the term phylogenetics.
2. What is phylogenetic analysis?
3. What is phylogenetic tree?
4. State about the rooted phylogenetic tree.
5. Explain about the rooted phylogenetic tree.
6. What is optimality criteria?
7. State the special tree types.
8. What is PHYLIP?
9. When ClustalW is used?
10. What is time complexity?

Long-Answer Questions

NOTES

1. Briefly discuss the concept of phylogenetics giving appropriate examples.
2. Explain the concept of phylogenetic analysis with the help of examples.
3. Discuss in detail about the phylogenetic tree and its significant properties.
4. Differentiate between the rooted phylogenetic tree and unrooted phylogenetic tree giving appropriate examples.
5. Elaborate the characteristic features of branching diagrams in phylogenetic analysis.
6. Briefly discuss how the phylogenetic relationships can be analysed and estimated using the optimality criteria.
7. How the construction of phylogenetic tree is done? Explain giving examples.
8. Briefly discuss the phylogenetic analysis using PHYLIP for rooted trees and unrooted trees.
9. Discuss how phylogenetic analysis is done using ClustalW.

14.7 FURTHER READINGS

Daniel, W.W. 2009. *Biostatistics: Foundation for Analysis in the Health Sciences,* 9th Edition. USA: Laurie Rosatone Publishers.

Agarwal, S.K. 2005. *Advanced Biophysics*. New Delhi: APH Publishing Corporations.

Mount, David W. 2004. *Bioinformatics: Sequence and Genome Analysis,* 2nd Edition. New York: Cold Spring Harbor Laboratory Press.

Leach, Andrew R. and Valerie J. Gillet. 2007. *An Introduction to Chemoinformatics*, Revised Edition. The Netherlands: Springer.

Gupta, C.B. and Vijay Gupta. 2004. *An Introduction to Statistical Methods,* 23rd Edition. New Delhi: Vikas Publishing House.

Pardo, Scott and Michael Pardo. 2018. *Statistical Methods for Field and Laboratory Studies in Behavioral Ecology,* 1st Edition. (Chapman and Hall/CRC). US: CRC Press.

Sokal, R. R. and F.J. Rohlf. 1969. *Biometry*: *The Principles and Practice of Statistics in Biological Research*. San Francisco: W.H. Freeman.

Pielou, E. C. 1984. *The Interpretation of Ecological Data: A Primer on Classification and Ordination,* 1st Edition. USA: Wiley-Interscience.

Rajaram, J. and Kuriacose, J.C.1993. *Electrochemical Methods Used in Electrode Kinetics, Reaction at Electrode Surface Kinetics and Mechanisms of Chemical Transformation,* 3rd Edition. New Delhi: Macmillan Publishers India Ltd.